Références scientifiques en ligne : folksonomies et activité des groupes

Evelyne Broudoux, Claire François, Besagni Dominique, Fabry Cécilia, Roussel Clotilde

Chapitre 23
Références scientifiques en ligne : folksonomies et activité des groupes

Dominique Besagni*, Evelyne Broudoux**
Cécilia Fabry*, Claire François*
Clotilde Roussel*

*INIST-CNRS
{prenom.nom}@inist.fr
http://recherche.inist.fr/
**DICEN - CNAM
evelyne.broudoux@cnam.fr
http://dicen.cnam.fr

Résumé. Cet article se propose de comparer l'activité des groupes de travail sur les serveurs d'enregistrements de références scientifiques en ligne : CiteUlike, Bibsonomy, Connotea et 2Collab et de repérer les domaines scientifiques représentés.

Abstract. This paper presents the results of a comparison between four online references manager services (CiteUlike, Bibsonomy, Connotea and 2Collab). We focus on the groups activity and try to detect the representation of the scientific domains.

1 Introduction

Cet article se propose de prolonger la réflexion initiée lors d'un Atelier Insi-Inist sur la notion de « folksonomies scientifiques » et présentée à Isko en 2009. Cette expression avait pour objectif de caractériser les traces de collecte ou de rassemblement de documents sur le web par ceux qui sont impliqués dans une recherche scientifique. Ces périodes de constitution de documentation se déroulent en parallèle aux recherches et se situent dans des périodes d'états de l'art, exploratoires, d'incertitudes et de questionnements.

Par folksonomie, nous entendons habituellement un ensemble de tags (ou mots-clés) produits par les internautes dans des moments de navigation, de consultation ou d'enregistrement de ressources sur le web dans un objectif de marquage en collectif. En réalité, la notion de « collectif » dans le cadre de l'utilisation d'outils de type « web2.0 » est toute relative car scénarisée par le formatage des accès en écriture/lecture d'espaces disques des serveurs. On peut d'ores et déjà retenir des similitudes conditionnant cette forme automatisée de « collectif », du point de vue des groupes :

- les groupes peuvent être publics ou privés et l'inscription peut être soumise à validation par le créateur du groupe ;
- les membres des groupes peuvent être anonymes ;
- l'alimentation en références des groupes par leurs membres est volontaire (et non automatique).

Caractéristiques générales des groupes :

- Groupe « test » avec un seul membre ;
- Groupe « bibliographie » avec un seul membre qui thésaurise de manière individuelle un ensemble de références correspondant à une thématique ;
- Groupe « informel » d'utilisateurs partageant volontairement et publiquement des références dans un domaine spécifique ;
- Groupe « formel » public ou privé, émanant d'un institut ou d'une association partageant des références spécialisées ;
- Groupe « formel structuré » reconnaissable par des acronymes (conférences, travaux dirigés d'étudiants, etc.).

1.1 Présentation des services ayant servi de base à l'étude

CiteUlike est un service en ligne de gestion et de partage de références bibliographiques créé par James Cameron en 2004 à l'Université de Manchester en Grande-Bretagne. Le service a été géré par Oversity Ltd créé pour l'occasion en 2006 et est sponsorisé par Springer depuis août 2008.

Cet outil collaboratif de gestion bibliographique comporte un volet éditorial puisqu'il permet de créer et gérer des listes d'articles (nommées bibliothèques) issus de 13508 périodiques présélectionnés[1]. Consultation de la table des matières d'une revue sélectionnée, extraction de références bibliographiques, importation et exportation sous différents formats de citation, permettent aux scientifiques de gérer listes de références et bibliographies.

Un volet « web 2.0 » comporte les fonctionnalités classiques de combinaison d'un espace personnel avec des fonctionnalités de partage et de mise en commun des ressources :

- tagging[2], évaluation et commentaires à partir d'articles issus des revues présélectionnées ou introduits de manière manuelle par les usagers, un système de recommandations propose automatiquement des références mises à jour dès que l'usager a un minimum de vingt articles enregistrés dans sa bibliothèque ;

- visualisation de ce que les autres lisent, comptage du nombre de personnes ayant tagué le même article, navigation sur les nuages de mots-clés classés par ordre quantitatif ;

- partage des ressources entre pairs, création de groupes thématiques, participation à des discussions, fonctionnalités de veille sur des mots-clés, auteurs, etc.

Les avantages du service sont qu'il extrait automatiquement les métadonnées de l'article sélectionné (auteurs, source, résumé, propositions de tags) et possède des fonctionnalités avancées de tri par facettes (date de postage, priorité de lecture, date

[1] Au 31 août 2010.
[2] Sur CiteUlike, on tague « en aveugle », au contraire des services qui proposent des mots-clés détectés comme « similaires » issus des documents ou créés par les autres usagers.

de publication, auteur, titre, type, journal, éditeur, ISSN, ISBN et clé BibTeX) dans la sélection des listes de références.

Aux fonctionnalités classiques de recherche d'articles, le service a rajouté depuis 2011 la classification scientifique des articles avec des sous-domaines détaillant leurs spécialités.

Lancé en 2004 par la revue Nature (Nature Publishing Group), **Connotea** est un outil de partage de signets dédié au monde académique qui propose de gérer, partager et découvrir des références bibliographiques par l'intermédiaire de tags et d'annotations. Comme pour CiteUlike, l'export des notices est possible sous quatre formats, bien qu'aucune distinction ne soit faite entre ressources publiées sur le web et publications scientifiques. L'extraction de références bibliographiques issues d'une dizaine de bases[3] est automatisée.

Un peu plus d'un an après son lancement, Connotea a opéré une distinction entre groupes publics et privés, et a donc rendu invisibles les groupes privés « puisque seuls les membres des groupes privés sont au courant de leur existence »[4]. Un projet d'unification sémantique des tags nommé « entity-describer » est tenté en 2007 mais il n'en reste plus de trace aujourd'hui.

Développé depuis 2006 par le Knowledge & Data Engineering Group de l'Université de Kassel, **Bibsonomy** est historiquement le premier service de partage de signets à proposer à l'usager d'opérer une hiérarchisation des tags à partir de la spécification de leurs relations (« supertag » et « subtag »). Les mots-clés généralistes deviennent alors des clusters à partir desquels on peut effectuer des recherches.

Comme CiteUlike et Connotea, Bibsonomy comporte toutes les facilités d'importation-exportation de notices bibliographiques mais celles-ci sont nettement plus étendues : plus d'une trentaine de formats différents dans l'exportation des références des publications.

Bibsonomy se différencie de CiteUlike par la possibilité de taguer n'importe quelle ressource du web et par sa notion élargie de « publication » : l'usager choisit au moment du référencement s'il poste un bookmark ou une publication dont pas moins de 10 types lui sont alors proposés.

Au contraire de CiteUlike, Bibsonomy se caractérise par un nombre réduit de groupes mais ceux-ci sont formels puisqu'ils émanent pour la plupart de laboratoires ou d'instituts scientifiques allemands et européens. L'onglet « Groups » réunissant les dénominations sous formes de liens externes pointant vers leurs sites web donne une indication de leurs orientations (sciences de l'information, des réseaux, informatique, apprentissages collaboratifs, etc.).

---

[3] Nature.com, PubMed, PubMed Central Science, PloS, BioMed Central, EPrints repositories, Highwire Press publications, Blackwell Synergy, Wiley Interscience, Scitation, arXiv.org, Smithsonian/NASA, Astrophysics Data System, Amazon, HubMed, D- Lib Magazine.

[4] « Groups can be either public or private — if private, only the group's members will be aware of its existence ». http://www.connotea.org/guide#groups

Dernier produit éditorial à être lancé sur le web2.0 scientifique, **2Collab** lancé par Elsevier fin 2007 apparaît au moment où la courbe d'adoption des innovations (selon Everett Rogers) marque le pas. Ce service en ligne qui permet d'enregistrer des références d'articles issus de Scopus ou de ScienceDirect n'attire pas un nombre suffisant d'usagers nécessaire à la collaboration massivement distribuée et est rapidement pris pour cible par les spammeurs[5].

1.2 Méthodologie

1.2.1 Collecte des données sur les sites

La collecte d'information sur les différents sites a été réalisée à l'aide de robots de recherche écrits en Perl avec le module Mechanize, module permettant à un programme de simuler les interactions d'un internaute sur son navigateur Web. La recherche commence à partir de la liste des groupes d'utilisateurs et le programme suit les liens conduisant aux pages propres des groupes, des utilisateurs, des tags et des références bibliographiques (ou des signets). Chaque page est analysée à l'aide d'un parseur HTML pour récupérer les informations pertinentes.

Dans le cas du site Bibsonomy qui possède une API d'interrogation plus efficace et plus rapide, le résultat est un fichier XML que l'on analyse avec un parseur XML, mais le principe reste le même.

1.2.2 Analyse thématique des groupes

L'étude de l'organisation thématique des groupes a été réalisée à l'aide de l'outil Neurodoc qui applique la méthode de classification des k-means axiales (Lelu & François (1992) ; Grivel & François (1995)). Cet outil étant conçu pour traiter des notices bibliographiques, nous avons généré de pseudo-notices où le nom du groupe correspond au titre du document et les tags à ses mots-clés.

La proximité entre deux groupes est donc fonction de la proportion de tags partagés, chaque groupe étant représenté par un vecteur de tags, et la proximité correspondant au produit scalaire des vecteurs. Les groupes sont donc rassemblés en classes en fonction de leur proximité thématique. Les classes sont ensuite projetées sur une carte selon une Analyse en Composantes Principales. Les distances entre classes sur la carte peuvent également être interprétées en fonction des proximités thématiques des classes.

[5] Il a cessé officiellement de fonctionner depuis le 15 avril 2011.

2 Résultats

2.1 Caractéristiques pour les 4 sites

	CiteUlike	Bibsonomy	Connotea	2Collab
Nombre de groupes	2871	171	397	5181
Dont nombre de groupes fermés	49%	-	24%	12%
Nombre total de membres*	10453	491	667	6414
Nombre total d'articles *	341498	79889	17162	51544
Nombre total de signets	/	69595	40680	/
Nombre total de tags différents	68522	34754		35057

TAB. 1 – Tableau général caractérisant les 4 sites

* Le nombre total de membres et d'articles est calculé à partir l'effectif de chaque groupe, les membres ou articles appartenant à plusieurs ne sont pas dédoublées.

CiteUlike et 2Collab regroupent beaucoup plus de membres que Bibsonomy et Connotea (facteur 10 au moins), cependant les membres de Bibsonomy sont particulièrement actifs car ce service regroupe autant d'articles et une aussi grande diversité de tags que Connotea et 2Collab.

2.2 Effectifs et dynamiques des groupes

	CiteUlike	Bibsonomy	Connotea*	2Collab
Effectif du groupe le plus important	99	47	43	44
Nombre de groupes ayant un seul membre (% du nombre total de groupes)	1434 49%	13 7%	193 48%	4655 90%
Nombre de groupes sans membre (% du nombre total de groupes)	-	85 50%	0	-

TAB. 2 – Membres des groupes : chiffres –clés

* on ne sait rien sur les groupes fermés, ils ne sont pas comptés dans ces statistiques

CiteUlike se distingue par le groupe avec le plus grand effectif. Pour les trois autres sites, quelque soit le nombre total de membres et de groupes, la taille maximale des groupes est sensiblement la même. La figure 1 montre que trois ont des effectifs très proches pour les 80 groupes les plus important. 2Collab se distingue par une grande proportion de groupes avec un seul membre (90% des groupes) tandis que, dans le cas de CiteUlike et de Connotea, seulement la moitié des groupes ne comprennent qu'un seul membre. Ces groupes qui ne comportent qu'un seul membre représentent soit des tests, soit une façon de créer des bibliothèques spécialisées. Bibsonomy comporte 85 groupes sans aucun membre mais ceux-ci sont cachés, ce qui n'empêche pas leur activité.

Plus précisément, la figure 1 montre que les groupes comprenant plus de 20 membres sont rares (moins de 5) pour tous les sites, excepté CiteUlike qui en compte environ 80.

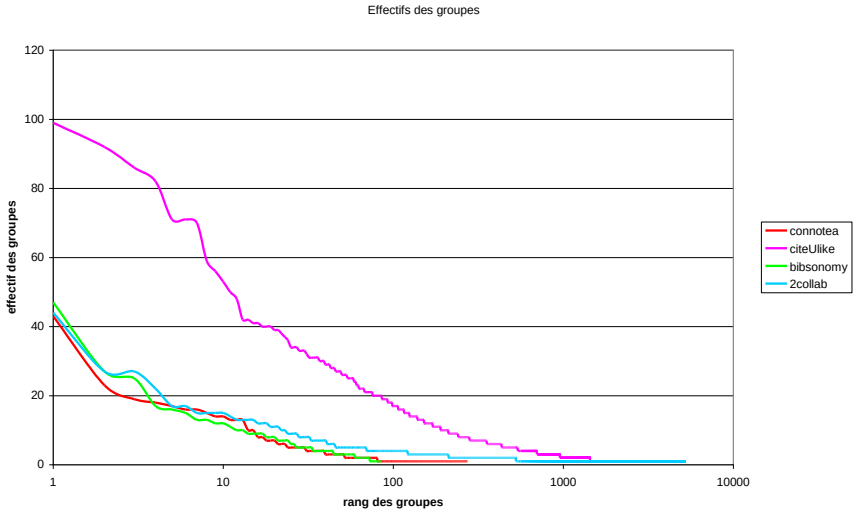

FIG. 1 – Distribution des groupes selon leurs effectifs en membres pour les 4 sites étudiés. Sur l'axe des abscisses, les groupes son triés par effectif décroissant.

La figure 2 permet d'analyser la dynamique des groupes en comparant le nombre d'ouvertures de groupe par an pour les 4 sites étudiées.

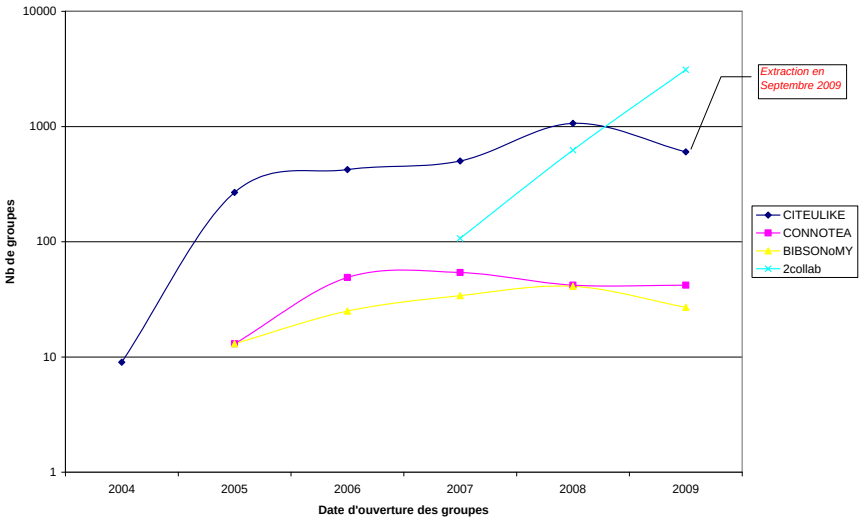

FIG. 2 – Nombre d'ouverture de groupes entre 2004 et 2009.

La courbe de CiteUlike montre que ce site est le plus ancien parmi les sites étudiés et que sont activité reste supérieure aux autres jusqu'en 2008, année où l'on observe le pic d'ouverture sur cette période, avec 1066 groupes créés. Dès 2009, 2Collab le site le plus récent présente le plus de groupes mais cette activité doit être relativisée par le fait que 90% des groupes ne possèdent qu'un seul membre. Pour l'ensemble de la période étudiée, sur CiteUlike sont ouverts tous les ans, au moins

10 fois plus de groupes que sur Connotea et sur Bibsonomy qui atteignent respectivement leur pic en 2007 et 2008. Remarquons que les chiffres d'ouverture commencent à baisser en 2009 pour CiteUlike et Bibsonomy, ce qui correspond à la baisse générale de l'usage de services de sauvegarde de signets en ligne (ex : Delicious) désertés pour Twitter qui rafle pour le moment la mise dans ce domaine.

2.3 Activité des groupes

Nous étudions l'activité des groupes au travers des articles et signets collectés par ces derniers.

	CiteUlike	Bibsonomy			Connotea*			2Collab
		Signets	Articles	Total	Signets	Articles	Total	
Nb articles ou signets maximal dans un groupe	27 741	5839	4778	10437	4068	1481	5400	9291
Nb groupes ayant un seul article ou signet	236	15	3	2	20	20	17	3153
(% du nombre total de groupes)	8%	9%	2%	1%	5%	5%	4%	61%
Nb groupes sans article ni signet	524	40	24	16	65	123	64	850
(% du nombre total de groupes)	18%	23%	14%	9%	16%	31%	16%	16%

TAB. 3 – Chiffres –clés des signets et des groupes

** on ne sait rien sur les groupes fermés, ils ne sont pas comptés dans ces statistiques*

CiteUlike est le site qui regroupe le maximum d'articles (tableau 1), ceci se retrouve également dans la taille maximale des groupes, avec un groupe rassemblant 27741 articles alors que pour Bibsonomy, 2Collab, et Connotea, les groupes les plus importants rassemblent environ 10000 et 5000 articles ou signets respectivement. 2Collab se distingue par le nombre très important de groupes avec un seul article (61% des groupes), ceci doit être mis en relation avec la très forte proportion de groupes avec un seul membre (90% des groupes). Le nombre de groupes vides (sans articles, ni signets) reste faible sur tous les sites, il se situe entre 9% et 18%.

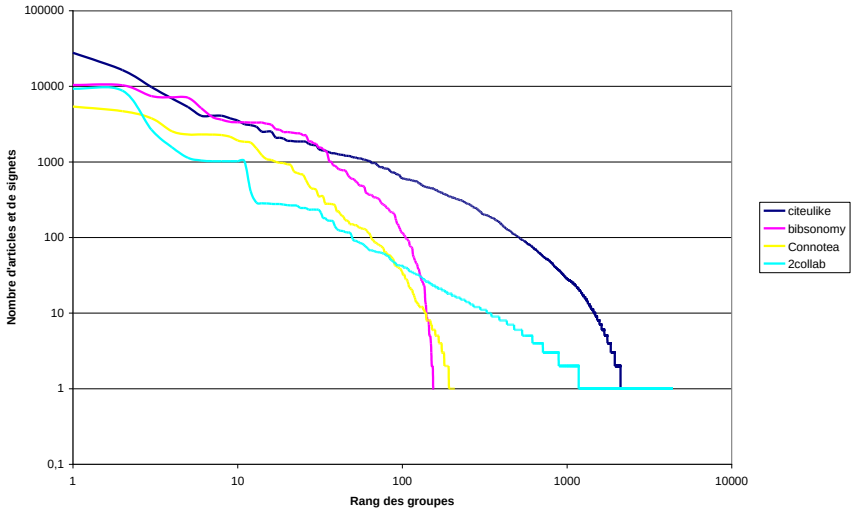

FIG. 3 – Distribution des groupes selon le nombre d'articles et de signets associés sur les 4 sites étudiés. Sur l'axe des abscisses, les groupes son triés par nombre d'articles et de signets décroissant.

La figure 4 permet une analyse plus détaillée de la distribution des articles dans les groupes pour les 4 sites étudiés. Ce mode de représentation permet de les comparer avec une distribution de type « Zpif ». Si les données respectaient exactement cette loi, les courbes auraient la forme de droites. La courbe qui s'en éloigne le plus est celle de « Bibsonomy », qui comprend comparativement peu de groupes, et également peu de groupes faiblement actifs (de 0 ou 1 articles ou signets), et où la proportion de groupes très actifs est importante : 21% des groupes ont plus de 1000 articles ou signets et 60% plus de 100.

Afin de mieux comprendre les usages du site CiteUlike, nous avons étudié plus particulièrement ses groupes les plus actifs. Le groupe rassemblant le maximum d'articles est le groupe C. elegans / Wormbase dont les deux premiers membres ont listé la totalité des articles en un peu plus d'un an, tous concernant les nématodes (ver ronds) et en particulier le 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).

Dix autres personnes se sont ajoutées comme membres à ce groupe après octobre 2009 (la dernière s'est ajoutée en avril 2011) alors que le dernier article a été posté en septembre 2009.

Avec 16363 articles, le groupe Computational Cognitive Neuroscience Lab (Understanding the Mind by Simulating the Brain) est le deuxième groupe le plus important de CiteUlike ayant engrangé des articles d'octobre 2005 à septembre 2008. Un seul usager apparaît avoir tagué la totalité des articles sauf un, les deux autres usagers s'étant inscrit en même temps pour le dernier article tagué.

Biodiversity_conservation (Biodiversity conservation, conservation biology, conservation Policy) avec ses 6438 articles et ses 73 inscrits arrive en troisième position. Il s'agit d'un groupe de juin 2005 qui reste actif tant en nouveaux membres qu'en articles postés.

Nous avons pu observer un phénomène particulièrement intéressant : une bibliothèque qui n'est plus mise à jour, partagée par 3 groupes dont le nombre de membres continue de croitre. Cette bibliothèque comprend 1862 articles indexés par 2938 tags, et le dernier article a été déposé en 2007. Les 10 tags les plus fréquents sont : bibtex-import, Species, Habitat, Spatial, Control, Management, Biodiversity, Conservation, Population et Diversity. Cette bibliothèque est partagée par les 3 groupes suivants :

- *Botany* : « taxonomie, évolution, physiologie, écologie, génétique », de 11 membres, créé en 2007, et dont le dernier utilisateur a été ajouté en août 2009,

- *Entomology* : « taxonomie, écologie, conservation, évolution, physiologie, génétique » de 10 membres, créé en 2007, et dont le dernier utilisateur a été ajouté en septembre 2009,

- *EarthEnvironmentalSciences* : « Recherches majeures sur les problèmes environnementaux, écologiques, socio-économiques, géographiques, de biodiversité. Du changement climatique terrestre aux études de porosité du sol. Des croutes microbiennes à l'étude de la forêt amazonienne. Des glaciers polaires au désert du Sahara », de 12 utilisateurs, créé en 2008, et dont le dernier utilisateur a été ajouté en août 2009.

Étant donné que la collecte des données a été réalisée en septembre 2009, nous pouvons considérer que ces groupes étaient encore actifs du point de vue des membres bien que la bibliothèque ait été constituée uniquement l'année de création du premier groupe (2007).

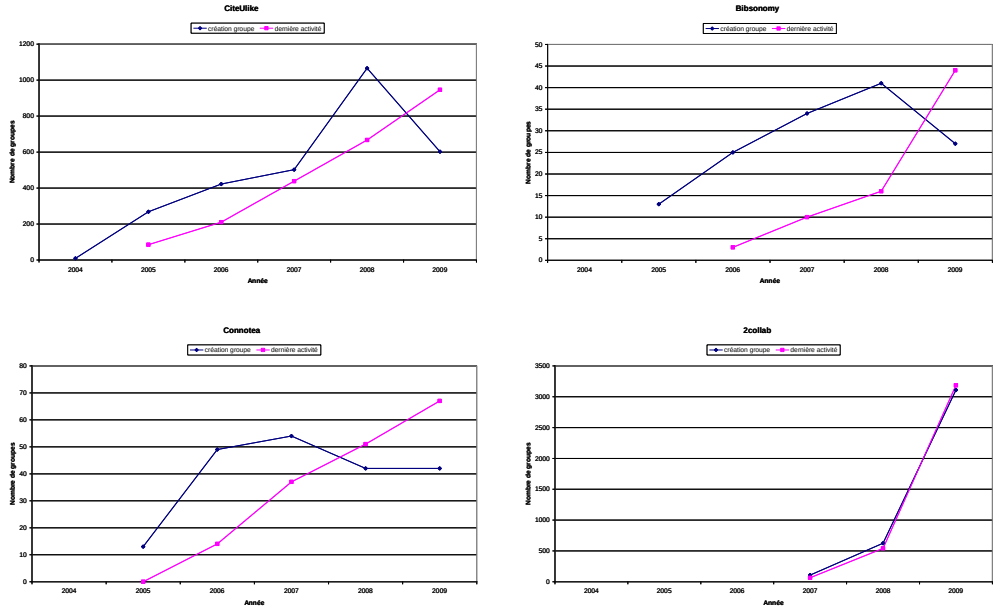

FIG. 4 – Comparaison entre les ouvertures et la dernière activité des groupes

Afin d'analyser la dynamique de l'activité des groupes, nous avons étudié le nombre de groupes ayant ajouté leur dernier article pour chaque année. Nous comparons les 4 sites étudiés selon ces informations.

Nous pouvons observer une dynamique de l'activité différente selon les sites. Concernant CiteUlike, nous observons une croissance régulière du nombre de groupes entre 2004 et 2007 avec un nombre de groupes cessant leur activité qui suit une croissance du même type. En 2008, nous observons une forte croissance de l'ouverture des groupes, alors que le nombre de groupes qui cessent leur activité ne croît pas dans les mêmes proportions, c'est donc une année de forte augmentation de l'activité en général. Etant donné que la collecte des données a eu lieu en 2009, le dernier point de cette carte correspond aux groupes encore actifs cette année 2009. La chute du nombre de création de groupes en 2009 doit être relativisée par le fait que la collecte des données a eu lieu en cours d'année.

Concernant Bibsonomy, après une croissance régulière des 2 courbes entre 2005 et 2008, nous observons une rupture entre 2008 et 2009 : chute du nombre d'ouverture de groupes et forte augmentation du nombre de groupe cessant leur activité. Cependant, la collecte des données s'étant déroulée en mars 2010, il est difficile de faire la part des choses entre un groupe faiblement actif et un groupe inactif depuis quelque mois.

Concernant Connotea, la rupture est observée déjà entre 2007 et 2008, avec une baisse puis stagnation du nombre de groupes ouverts, et une continuité de l'augmentation du nombre de groupe cessant leur activité.

Concernant 2collab, les deux courbes sont tellement proches, que nous pouvons en déduire que la durée d'activité des groupes ne dépasse pas l'année.

2.4 Indexation par les tags

	CiteUlike	**Bibsonomy**	**Connotea**	**2Collab**
Fréquence maximale des tags	47 178	18207	-	16155
Nb de tags de fréquence 1 (% du nombre total de tags)	26 668 39%	13048 38%	-	16842 48%
Nb de tags dans un seul groupe (% du nombre total de tags)	43 820 64%	18388 53%	-	19469 56%

TAB. 4 – Chiffres –clés de la fréquence des tags

Les tags ont une répartition dans les documents et dans les groupes également de type « loi de Zipf ». La proportion de tags de fréquence 1 est très élevée principalement pour 2Collab. La différence de proportion entre le nombre de tags de fréquence et le nombre de tags dans un seul groupe est plus élevée laissant présager des groupes thématiques.

Pour les différents services, l'analyse des tags les plus fréquents dans les groupes donne les résultats suivants :

- pour CiteUlike, ces tags communs aux différents groupes sont très généraux (model, design, theory, system, method, development, evaluation, research) associés à un vocabulaire informatique et mathématique (web, software, simulation, statistic) ;

- pour Bibsonomy, ce sont des catégories liées au web d'où émerge un profil informationnel (ontology, semantic, information, classification, evaluation) et des aspects sociaux (community, collaboration, social) ;
- pour 2Collab, apparaît un lexique lié au droit et à la justice.

En ce qui concerne les tags les plus fréquemment utilisés par article de CiteULike, apparaissent des lexiques spécialisés : c_elegans, nematode, elegans, caenorhabditis_elegans, wormbase, mdb (signication : Metalloprotein Database and Browser) employés par la biologie moléculaire et climate_change, evolution, forest, appartenant aux sciences de la terre, ainsi que digital-library aux sciences de l'information.

Pour Bibsonomy, les mots-clés apparaissent plus généraux (web, Learning) ou se spécialisant en sciences de l'information (folksonomy, semantic, social, tagging, design, ontology, clustering) ou bien à caractère auto-référentiel (my-own).

Pour 2Collab, le vocabulaire est lié à un individu ou une institution (petrauniv, petrathesis, undergrathes) et dans les divers tags, il est facile de reconnaître le spam (FREE theory test, car theory test, mock theory test, driving test, etc.).

La comparaison n'a été possible qu'entre CiteUlike et Bibsonomy, les chiffres de Connotea étant inaccessibles et ceux de 2Collab faussés par le spam. On remarque qu'au sein de Bibsonomy, les tags les plus fréquents dans les groupes sont également les plus fréquents dans les articles. Sur CiteUlike, les tags les plus fréquents dans les articles sont spécialisés et sont différents de ceux largement partagés par les groupes qui possèdent un caractère plus général.

3 Répartition des groupes par domaine pour CiteUlike

Nous présenterons les résultats de classification des groupes en fonction des tags utilisés. Pour cela, nous avons appliqué l'outil Neurodoc sur les groupes et l'index des tags associés à ces derniers en conservant les tags présents dans au moins 5 groupes. Ce filtrage a sélectionné 2216 groupes. Les intitulés de classes correspondent aux tags de poids le plus élevé pour chaque classe.

Nous étudions ci-dessous comment les tags permettent de représenter la thématique des groupes et comment ils permettent ou non d'organiser les groupes selon leurs domaines scientifiques.

3.1 Interprétation des vingt classes obtenues

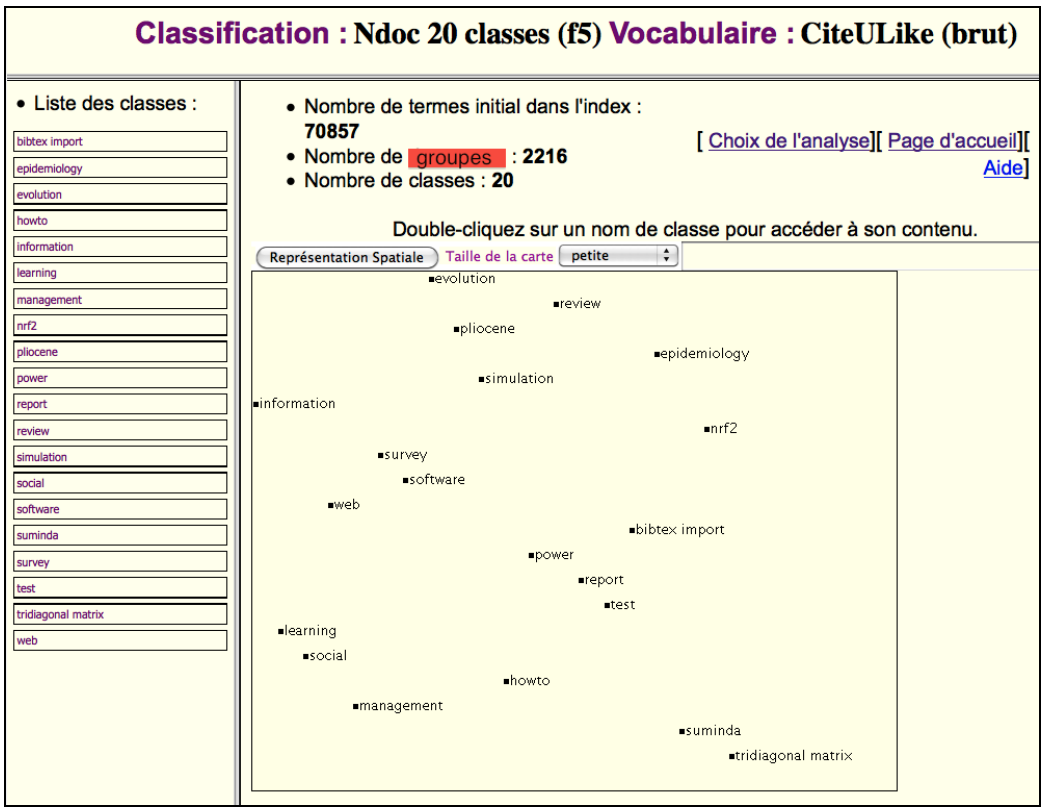

FIG 5 – Liste des vingt classes et représentation par rang.

Les intitulés des classes se répartissent de la façon suivante :
- 1 intitulé est généré lors d'une importation automatique de bibliographies (bibtex import) ;
- 5 intitulés sur 20 représentent une caractérisation de l'information stockée ou sont des documents pour l'action : howto, report, review, survey, test ;
- 5 intitulés sont représentatifs d'une méthodologie ou caractéristique d'une science : epidemiology, evolution, pliocene, simulation, tridiagonal matrix ;
- 6 intitulés sont représentatifs de grandes catégories : learning, management, power, social, software, web ;
- 1 intitulé est à la fois représentatif d'une science tout en étant une catégorie générale : information ;
- 1 intitulé est un acronyme (nrf2) ;
- 1 intitulé apparaît ne rien signifier (suminda).

Nous avons relevé les premiers tags ramenés par Neurodoc pour chaque classe et avons complété les observations en consultant directement sur CiteUlike les caractéristiques des groupes concernés lorsqu'elles étaient disponibles.

3.1.1 Domaine : sciences de la vie

La classe « Evolution » est la classe la plus active avec 777 tags et 241 groupes pour la plupart formellement identifiés et des articles postés régulièrement dans ce champ de recherche qui a la particularité de faire converger plusieurs disciplines autour de la bio-informatique.

Les deux premiers sous-clusters de groupes indexés par les tags « evolution » et « rna » représentent bien l'essentiel de la classe avec les mêmes groupes de tête :

- *BioinfoCIPF* (Département bioinformatique du CIPF) : 10 usagers, 1039 articles ;
- *Bioinformatics* (Analyse et modélisation des données de biologie moléculaire) : 173 usagers, 4285 articles ;
- *EisenLab* (Groupe Michael Eisen's de UC Berkeley) : 19 usagers, 1115 articles ;
- *Bioinplant* (Bioinplant Lab de l'Université Zhejiang) : 1 usager, 69 articles ;
- *microRNA* (articles sur les microRNAs (computationel, wetlab, clinique)) : 41 usagers, 1836 articles.

Il est remarquable que la classe « Evolution » qui émerge en premier concentre des disciplines différentes sur un même objet de recherche (l'adn, le génome) avec la bio-informatique comme élément fédérateur. Ainsi, l'évolution est ici vue essentiellement comme biologique.

3.1.2 Domaine : Sciences de la terre

La classe « Pliocene », dont le terme signifie une époque précise sur l'échelle des temps géologiques, est représentative de la place prise par la Géologie et les Sciences de la terre sur CiteULike. Elle se compose de 3117 tags et de 85 groupes.

Le premier sous-cluster de groupes indexés par le tag « Pliocene » rassemble 5 groupes : *biodiversity_conservation, climate_change, Climate Change, Tectonics_and_geomorphology, palm library (Arecaceae, Palmae).*

- *biodiversity_conservation* (conservation de la biodiversité, conservation biologique, politique de conservation) se compose de 72 usagers ayant tagué 6434 articles. Le groupe créé en avril 2005 continue son activité car le dernier article posté date du 13 avril 2011 et il continue de gagner en membres ;
- *climate_change* (articles relatifs au changement climatique) se compose de 26 usagers ayant tagué 1160 articles ;
- *Climate_Change* n'existe plus ou a transformé son titre ;
- *Tectonics_and_geomorphology* (Déformation continentale et tectonique, tectonique active, géomorphologie de la tectonique, climat et tectonique, géodynamique, tectonique des orogènes collisionnels) : 1 usager ayant tagué 250 articles (février 2005 - février 2007) ;

- *palm library* (Arecaceae, Palmae) (Pour une plus complète (et peut-être plus exhaustive) bibliographie sur les palmiers. Toute nouvelle référence sur les palmiers est bienvenue. Systématique, anatomie, paléobotanique, archéobotanique, ethnobotanique, morphologie, écologie, évolution...) : 1 usager, 481 articles ;

Le deuxième sous-cluster de groupes indexés par le tag « Miocene » comprend 4 groupes appartenant au tag précédent (*Pliocene*) auxquels s'ajoutent *Global_biodiversity_model* et [*BrahmaTWinn*] -> export to iw, que l'on ne retrouve plus dans la base actuelle.

Le troisième sous-cluster de groupes indexés par le tag « Landsat » composé de 13 groupes se distingue nettement des deux précédents par des orientations plus généralisantes en termes taxonomiques :

- *Botany* (décrit en § 2.3) : 13 usagers, 1862 articles (mai 2007 – juin 2007) ;

- *Zoology* (taxonomie, paléontologie, physiologie, évolution, anatomie, taxonomie, écologie, génétique) : 14 usagers, 1879 articles (juin 2007 – mars 2010) ;

- *Entomology* (décrit en § 2.3): 13 usagers, 1868 articles (mai 2007 - décembre 2010) ;

- *Ecology* (écologie de population, écologie de communité, écologie théorique) : 45 usagers, 1919 articles (mai 2007 - octobre 2010) ;

- *EarthEnvironmentalSciences* (décrit en § 2.3) : 17 usagers, 1871 articles (mai 2007 - mai 2010) ;

- *Computational_Systems_Biology* (Biologie des systèmes computationnels) : 47 usagers, 2090 articles (octobre 2005 - avril 2011) ;

- *GrassBase* (Usagers de GrassBase - the Online World Grass Flora) : 2 usagers, 2108 articles (mars 2006 – septembre 2007) ;

- *biodiversity_conservation* (conservation de la biodiversité, conservation biologique, politique de conservation) : 72 usagers, 6434 articles, (avril 2005 – avril 2011) ;

- *Tree Species Study* (de la vulnérabilité des espèces arboricoles du Canada aux impacts du changement climatique, biologique et de la capacité humaine d'adaptation) : 4 usagers, 4072 articles ;

- *Landscape Ecology & Conservation* : 14 membres, 56 articles (novembre 2007 à mai 2010) ;

- *Tectonics_and_geomorphology* (décrit en 3.1.2) ;

- *Global_biodiversity_model* et *Remote Sensing & GIS* sont deux groupes dont il ne subsiste que le titre.

Notons que *Botany*, *Zoology*, *Entomology*, *Ecology*, *EarthEnvironmentalSciences* ont été créés par le même utilisateur : malaeng qui poste et importe quasi-systématiquement dans 8 groupes en même temps (*biodiversity_conservation*, *Computational_Systems,_Biology*, *GrassBase*, *Entomology*) et se constitue ainsi une bibliothèque de 1862 articles tagués.

L'exploration de la classe « Pliocene » dégage le profil suivant : les groupes taguent de manière intensive des milliers de références ayant traits aux sciences environnementales, aux sciences de la terre, sur des périodes n'excédant pas deux ou trois ans. Composés d'un seul membre, d'une dizaine ou de soixante à quatre-vingt membres, les groupes sont informels et ont pour la plupart cessé toute activité. Les groupes continuant leur activité de taguage sont ceux qui comportent le plus de membres. Aucun groupe n'est formel (pas d'institution revendiquée).

Concernant le domaine scientifique représenté, contrairement à la classe précédente où l'on pouvait remarquer une focalisation convergente sur un aspect spécifique de l'évolution, il est ici très diversifié et divergent. En effet sous le vocable « pliocène » sont rassemblées les sciences de la terre dans leur totalité avec la géologie en tête. Cependant, les centres d'intérêts spécifiques qui apparaissent comme la conservation de la biodiversité, les sciences environnementales, l'impact du changement climatique sur la géographie, la socio-économie ou les capacités d'adaptation de la biosphère indiquent des préoccupations sociétales prises dans l'actualité.

3.1.3 Domaine : Médecine

La classe « Epidémiology » de 177 tags et de 43 groupes représente deux dominantes d'intérêts. On y distingue des items médicaux (geriatrics, endocrinology, thyroïd, renal, vascular, healthcare, fiber, adrenal) de ceux plus spécifiquement liés à l'environnement et au traitement de l'eau (hydrologie, groundwater, infiltration, climate change, microfluidics, sedimentation, geomorphology). Ceci répond parfaitement à la définition de l'épidémiologie qui concerne la santé publique avec l'étude des facteurs influant sur la santé et les maladies des populations humaines.

Un seul des huit groupes du sous-cluster de groupes indexés par le tag « epidemiology » est identifiable (*SepNet*), émanant de la German sepsis society. Deux groupes informels font appel à une méthodologie spécifique : *Randomized-clinical-trials-review* (Révision d'articles RCT actuels) et *Randomized-clinical-trials-methodology* (Révision d'articles sur le design, l'implémentation, and l'analyse des RCT). Deux autres groupes informels s'adressent explicitement à des chercheurs spécialisés : *Mathematical epidemiology* (pour les biologistes et les mathématiciens dont la recherche concerne les modèles mathématiques des maladies) et *Epidemiology - pharm.epi.* (pour les épidemiologistes travaillant dans les branches pharmaceutiques/biotech/médicales des entreprises équipementières).

Nous pouvons rapprocher de cette classe celle intitulée nrf2 composée de 1459 tags et 98 groupes dont 70 sont des émanations de l'HEIRS[6], une organisation 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.

Nrf2 est une protéine de liaison de l'ADN, dont le système antioxydant peut être affecté négativement par les métaux lourds. D'après HEIRS, les américains de souche et leurs descendants connaissent un risque accru - à cause de leur polymor-

[6] Health Education Information and Resource Services (HEIRS)

phisme hérité - d'être touché par des problèmes de santé dus à la détérioration de cette protéine par les dégradations environnementales[7].

Les groupes de la classe « Epidemiology » comportent peu d'articles dans leurs bibliothèques, au contraire du cluster « nrf2 » qui avec ses 70 groupes authentifiés HEIRS a engrangé un total de 15547 articles soit 4,5 % de la base CiteUlike. Le sous-titrage précis de chaque groupe laisse supposer que les articles sont suffisamment différents pour éviter d'être tagués plusieurs fois. Les groupes dans un cluster comme dans l'autre ont peu de membres (<9).

La classe « Epidemiology » qui là aussi peut être considéré comme un attracteur de différents domaines scientifiques trahit des préoccupations sociétales concernant le développement de maladies liées à des facteurs environnementaux.

3.1.4 Domaine : Sciences de l'information

Avec 543 tags et 560 groupes, la classe « Information » révèle bien l'écartement du champ entre des composantes liées à l'informatique, au web et à l'information et communication (sous-clusters de groupes indexés par les tags: ontology, folksonomy, tagging, usability, trust, sociology, data mining, architecture, retrieval, etc.).

A la tête des 179 groupes du sous-cluster de groupes indexés par le tag « information » se trouvent le groupe CSWC (pour personnes intéressées par le Computer-supported cooperative work), 55 membres, 567 articles postés entre janvier 2005 et septembre 2009. Notons que la créatrice du groupe Sylvie Noël l'est également de celui sur l'écriture collaborative (201 articles).

Viennent ensuite quatre groupes créés par le même utilisateur et qui contiennent pour les trois premiers le même nombre d'articles (970) postés entre novembre 2005 et mai 2007 et un seul membre.

- *vds-arg* (Kaleidoscope VDS en argumentation). Il s'agit d'un groupe émanant d'une école doctorale virtuelle (Kaleidoscope VDS) ;
- *dtl* (technologie du design et groupe d'études at http://www.lkl.ac.uk) ;
- *eni* (intelligence émergente de réseaux) ;
- *mathgamespatterns* (Groupe de recherche des modèles de jeux mathématiques, London Knowledge Lab and partners) - 5 membres, 1025 articles postés entre février 2005 – mai 2007.

Les sous-clusters suivants de groupes indexés par les tags (design avec 170 groupes), technology (131), theory : fréquence (173) contiennent pratiquement les mêmes groupes que le cluster information.

Au rang 17 le groupe « *philosophy of information* » (Recherche sur la philosophie de l'information) en activité avec 228 usagers et 1674 articles reflète bien la diversité des approches.

[7] http://heirsonline.blogspot.com/2010/11/nrf2-heavy-metals-and-mining-real-toxic.html

3.1.5 Classe Suminda et l'effet spam

En janvier 2006, l'usager Sirinath[8] poste un unique article fictif intitulé « My general bookmarks » et le tague avec cinq items : dharmasena, salpitikorala, sirinath, suminda. Cet unique article est ensuite enregistré systématiquement dans 16 groupes créés pour l'occasion et 4 autres existants (*GeneticAlgorithms, CompArch, Compilers* et *AI*).

Si Sirinath est resté l'unique membre des groupes : *Buddhism, vipassana, NeuroEvolution, CIS, ExpertSystems, PersonalDevelopment, StressManagement,* d'autres groupes ont recueilli des membres supplémentaires mais ont tagué moins de cinq articles : *Bayesian, FuzzyLogic, GeneticAlgorithms*

Les groupes suivants ont eu une réelle activité : *AI* (fin 2008), *DSS* (mi-2009), *GraphTheory* (fin 2009), *IntelligentAgents* (début 2009), *KnowledgeEngineering* (fin 2006), *0R* (début 2008), *DSS* (fin 2009). Quatre groupes peuvent être considérés comme encore actifs comme *CompArch, Compilers, NeuralNetworks, AIM,* puisque des signets y ont été postés avant ou pendant la rédaction de l'article.

Cet exemple peut servir à illustrer deux tendances : la propension à tester le système (ici est démontrée sa vulnérabilité et son détournement en tant qu'instrument de référencement) la faculté de régénérescence des groupes puisqu'un groupe peut se constituer à partir d'un article inexistant et attirer plus tard de nouveaux membres qui enregistreront des articles réels (par exemple, cinq ans après pour *AIM*).

3.1.6 Classe Tridiagonal matrix

Composée de 49 tags et de 24 groupes, la classe tridiagonal matrix est la plus petite de l'ensemble étudié. Selon Wikipédia, une matrice tridiagonale est une matrice dont tous les coefficients qui ne sont ni sur la diagonale principale, ni sur la diagonale juste au-dessus, ni sur la diagonale juste en-dessous, sont nuls. Il s'agit donc d'une expression issue des mathématiques.

Le premier sous-cluster de groupes indexés par le tag « tridiagonal matrix » se compose de 8 groupes : *Harmonic_Balance, Circadian_Rhythm, Random_Walks, Delay_Differential_Equations, Trajectory_Generation, Networked_Control_Systems, Autonomous_Vehicles, Machining_shear_bands.*

Le groupe « Harmonic_Balance » (La méthode d'équilibrage harmonique pour les systèmes non linéaires) comporte un seul membre ayant tagué 15 articles sur les équations différentielles entre août 2005 et avril 2007.

Les sous-clusters de groupes indexés par les tags « poincare lindstedt », « perturbation method », « impulsive », « génération funtion », « fare », « center manifold », comportent quasiment les mêmes groupes et concernent bien la thématique de la classe.

La classe « tridiagonal matrix » fait émerger l'intérêt pratique pour des applications réalisées à partir d'équations mathématiques (ex : génération de trajectoires,

[8] On identifie cet usager sur Twitter.

véhicules autonomes, modélisations de déformations de matériaux thermoplastiques, etc.).

D'une manière générale on peut retenir de cette étude que les groupes couvrent des thématiques qui relient des secteurs scientifiques à l'actualité sociétale vive. Les références partagées sont des indices de lectures correspondant à des préoccupations concrètes de la vie en commun.

3.2 Logiques de membres et logiques de groupes

Les usagers créateurs de groupes dans la base CiteUlike sont les plus actifs à taguer dans les groupes, ils en sont les animateurs et sont les premiers aussi bien que les derniers à poster des signets. Ils créent fréquemment d'autres groupes et enregistrent de manière systématique leurs références dans plusieurs groupes en même temps. C'est d'ailleurs une caractéristique remarquable que les groupes remontant en tête dans les clusters sont aussi ceux qui sont animés par les mêmes utilisateurs.

Les utilisateurs renseignent rarement leur profil et restent la plupart du temps anonymes, sauf lorsqu'ils ouvrent et gèrent des groupes, leur profil alors est plus souvent renseigné.

Peu de groupes apparaissent jouer en collectif, c'est-à-dire taguent de manière régulière dans un pot commun ; ce qui est au contraire lisible ce sont des logiques « individuelles » dont la vie des groupes reste dépendante.

4 Conclusion

Les grands domaines scientifiques représentés dans CiteUlike sont les sciences de la vie et de la terre, la médecine et les sciences de l'information et de la communication. Un domaine bien identifié se détache en termes d'activité, celui des sciences de la vie et en particulier celui de la bio-informatique. Il n'a pas été possible de distinguer les logiques de recherche de celles de publication et de la documentation sauf lorsque l'intitulé des groupes le permettait (ex : HEIRS), mais le fait qu'une ou deux personnes taguent pendant des périodes de quelques mois à un ou deux ans laisse supposer que c'est l'activité de documentation qui prime sans doute au service d'une équipe et/ou d'un projet.

On distingue deux figures de concentration thématique de références : celle convergeant vers un sujet unique engageant plusieurs disciplines et celle plus horizontale agglomérant les multiples composants d'un même domaine. Ces références traduisent aussi des préoccupations sociales fortes reliées aux domaines scientifiques représentés ; il est possible de les considérer comme des tendances de lecture.

L'usage de ces services en ligne suit des logiques de test (nombre important de groupes sans membres ou sans articles) et semble être guidé par la courbe du cycle de vie des innovations (déclin de l'activité à partir de la fin 2009).

Les limites de l'étude sont d'origines diverses : la collecte des données ne s'est pas déroulée dans le même temps (2009 et 2010), une mise à jour serait nécessaire pour parfaire les résultats. Des biais existent à travers l'importation automatique de

références susceptible d'être exploitée par le spam et les bases doivent être nettoyées régulièrement. On peut se demander si les pratiques elles-mêmes sont encore d'actualité ou si d'autres types d'outils ne sont pas venus les accaparer.

Ont été utilisées ici les méthodes d'analyse bibliométriques réalisées habituellement sur les documents indexés. Elles ont été adaptées aux documents tagués rassemblés dans les bibliothèques des groupes mais il resterait à analyser les réseaux des utilisateurs : comment observer certains groupes et les acteurs de ces groupes. Des réseaux bipartis resteraient à étudier : le réseau des documents dans les bibliothèques des groupes et les réseaux des utilisateurs participants ou non à différents groupes.

Références

[BOG, 08] Bogers T., van der Bosch A. « Recommending Scientific Articles Using CiteULike ». RecSys'08, Lausanne, Switzerland, 2008.

[BRO, 10] Broudoux E., François C. Besagni D., Fabri C. « Étude comparative du partage de références scientifiques (CiteUlike, Bibsonomy, 2Collab, Connotea) ». Séminaire Folksonomies et Tagging, Dicen – Cnam, 2010. (Document en ligne sur http://w1p.fr/21660).

[CAP, 07] Capocci A., Caldarelli G. « Folksonomies and clustering in the collaborative system CiteULike ». arXiv :0710.2835v2. 2007.

[DUN, 08] Duncan H., Pettifer S., Kell D. « Defrosting the digital Library : bibliographic tools for the next generation web ». *Plos Computational Biology*. Vol. 4. Issue 10, 2008.

[GRI, 95] Grivel G., François C. « Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique ». *Solaris* n°2, 1995. Document en ligne sur http://biblio-fr.info.unicaen.fr/bnum/jelec /Solaris/d02/2grivel.html

[KIP, 11] Kipp M. « User, Author and Professional Indexing in context : an exploration of tagging practices on CiteUlike ». *CJILS/RCSIB* 35, n°1, 2011.

[LEL, 92] Lelu A., François C. « Information retrieval based on a Information retrieval based on a unsupervised extraction of thematic fussy clusters », Neuro-Nîmes 92 : Les réseaux neuro-mimétiques et leurs applications, 2-6 novembre 1992, Nîmes, France.

[MUN, 07] Munk T., Mork K. Folksonomies, Tagging Communities, and Tagging Strategies-An Empirical Study. *Knowledge organization.* Allemagne. Vol. 34 ; N°3 ; pp. 115-127, 2007.

[OLD, 09] Oldenburg S. « Comparison of social classification systems in a heterogeneous environment ». Webist 2008, *LNBIP* 18, pp. 333-346, 2009.

[STO, 07] Stock W. (2007). « Folksonomies and science communication ». *Information Services & Uses* 27. IOS Press. pp. 97-103.