Le temps des SIC

Gabriel Gallezot, Marty Emmanuel

Version auteurs et en couleur du texte :
GALLEZOT, Gabriel et MARTY, Emmanuel, 2017. Le temps des SIC. In : MIÈGE, Bernard, PELISSIER, Nicolas et DOMENGET, Jean-Claude, Temps et temporalités en information-communication: Des concepts aux méthodes. Paris : L'Harmattan. pp. 27-44. ISBN 978-2-14-004014-6.

Le temps des SIC

"SIC" time

Gallezot Gabriel (1, 2) & Marty Emmanuel (1)
(1) Université Côte d'Azur, //TransitionS, France
(2) Urfist (Unité régionale de Formation à l'Information Scientifique et Technique)

gabriel.gallezot@unice.fr
emmanuel.marty@unice.fr

Mots clés : lexicométrie, analyse lexicale, fouille de texte, épistémologie,
Keywords : lexicometry, lexicon, text mining, epistemology

Résumé :
Pour rendre compte du temps des Sciences de l'Information et de la Communication (SIC), nous avons choisi d'analyser le lexique des chercheurs. Notre étude s'appuie sur les textes librement déposés par les auteurs sur la plateforme HAL/@sic. La fouille de texte s'effectue par une série d'analyses lexicométriques afin de répondre à deux objectifs : appréhender les notions liées au temps dans les recherches en SIC, d'une part, d'autre part rendre compte de l'évolution dans le temps des champs et questions de recherche en SIC.

Abstract :
To understand the notions of temporality in Information and Communication Sciences (ICS), we choose to analyze lexicon of researchers. Our study is based on texts freely submitted by authors on platform HAL / @ sic. Text mining is carried out by several lexicometric analysis, with two aims : to understand the concepts related to "time" in ICS' research and identify changes over time in ICS' researchers' fields and questions.

Le temps des SIC
“SIC” time

Gallezot Gabriel & **Marty Emmanuel**

1 - Introduction

Les SIC (Sciences de l'Information et de la Communication), association des Sciences de l’Information (SI) et des Sciences la Communication (SC) sont une spécificité française. Bien que les sous-champs des SIC aient leurs équivalents internationaux, la discipline est continuellement à la recherche de son identité. C’est certes le cas de toutes les disciplines, mais l'adjonction par nécessité « académique » des SI et SC questionne peut-être plus profondément encore notre communauté sur son épistémologie[1]. Depuis les années 1970, les SIC sont qualifiées de pluridisplinaires, transdisciplinaires, interdisciplinaires, comme si le terme de discipline ne pouvait convenir. Face aux questionnements sur la cohésion des SIC, la formule usuelle est celle d’une unité plurielle.

Cette étude est exploratoire. Elle se donne pour objet de mettre en lumière l’approche des SIC sur les notions liées au temps, question d’ordre épistémologique, ainsi que les éventuelles évolutions de ces approches dans une perspective diachronique. Le thème de ce XXe congrès de la SFSIC nous donne l’occasion de nous pencher sur le “temps des SIC” avec un double objectif : repérer, caractériser et contextualiser les notions et concepts relatifs au temps et aux temporalités qui ont été travaillés en SIC, d’une part[2] ; d’autre part appréhender la structuration des SIC et les évolutions de ses champs, objets et/ou terrains de recherche dans le temps[3]. Pour le dire autrement il s’agit de rendre compte de notre unité plurielle sur une période donnée et à travers le thème du temps. Deux angles d’approche capables de mettre en lumière le temps des SIC.

Notre culture commune se constitue en partie par les textes que nous produisons depuis la création de la discpline. Ces artefacts synchroniques marquent le temps autant qu’ils sont marqués par lui, et constituent les traces de notre développement. C’est la sédimentation de

1 Voir Jean Meyriat lors de la création du Comité des Sic le 25 février 1972 : " le mot plus concret d'information précise un peu la notion vague de communication ; ce couplage permet en même temps de servir les intérêts de plusieurs groupes distincts de spécialistes, sans prendre une position définitive sur l'épistémologie du domaine" : Boure, R. (dir.), Quelle histoire pour les Sic, in Les origines des sciences de l’information et de la communication, regards croisés. Septentrion, Villeneuves d’Ascq, 2002. (p.10)

2 Voir paragraphes 3.1 et 3.2

3 Voir paragraphes 3.1 et 3.3

ces traces que nous allons interroger pour tenter de définir les approches et usages du temps en contexte et effectuer ainsi une certaine archéologie des dépôts des chercheurs en SIC. Foucault (1969) pourrait dire : chercher "à définir dans le tissu documentaire lui-même des unités, des ensembles, des séries, des rapports" (p.14) et montrer que "l'histoire d'un concept n'est pas, en tout et pour tout, celle de son affinement progressif, de sa rationalité continûment croissante, de son gradient d'abstraction, mais celle de ses divers champs de constitution et de validité, celle de ses règles successives d'usage, des milieux théoriques multiples où s'est poursuivie et achevée son élaboration" (p.11).

2 - Outils, méthodes et corpus

Appréhender la diversité des domaines et objets qui composent notre unité est une préoccupation qui suscite régulièrement des travaux (Lancien T. et al., 2001 ; Boure R., 2002 ; Cardy H., Froissart P., 2002 ; Jeanneret y. et Ollivier B., 2004 ; Dumas P et al., 2006) aux méthodes très diverses : bibliométrie, scientométrie, analyses de réseaux et analyse qualitative, etc. La statistique textuelle semble étrangement n'avoir pas encore été pleinement mobilisée à cet égard. Dans le cadre de notre étude, basée sur la fouille et la caractérisation des artefacts textuels, cette méthode d'analyse nous paraît opérationnelle et adaptée. Notre démarche consiste donc à déterminer "le temps des SIC" par une méthode lexicométrique, à l'aide du logiciel Iramuteq[4] (Ratinaud & Dejean, 2009), qui implémente notamment la classification descendante hiérarchique développée par Reinert (1983).

Notre étude s'appuie sur les textes librement déposés par les auteurs sur la plateforme HAL plus précisément sur une de ses instances dédiées à notre discipline : @rchiveSIC. A l'aide de l'API de HAL[5], nous avons extrait les notices (titre, résumé, date publication, date de dépôt, et d'autres éléments qui ne servent pas directement l'analyse) de tous les dépôts en date du 17 mars 2016 (références seules ou accompagnées du texte intégral). Notre corpus initial est plus exactement constitué du titre et du résumé[6] de 4672 notices entre 2002 et 2016, soit sur les 15 dernières années, avec des dates de publication entre 1977 et 2016. Cependant la quasi

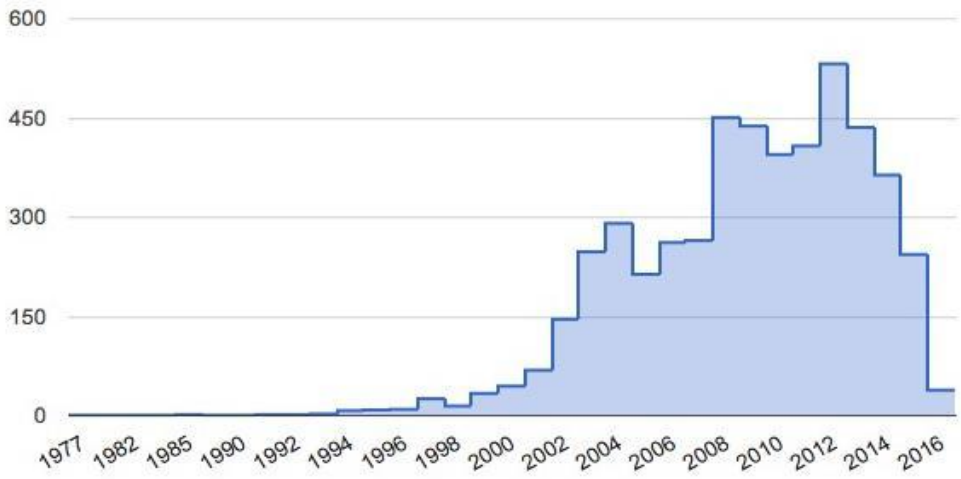
4 Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires . Développé sous licence GNU/GPL par Pierre Ratinaud, Lerass, Université Toulouse 3. http://www.iramuteq.org/

5 Application Programming Interface de HAL : http://api.archives-ouvertes.fr/docs

6 Le choix de travailler sur les titres et résumés et non sur le texte intégral a été motivé par le fait qu'environ la moitié des dépôts comprenaient le texte intégral, ce qui nous aurait privés d'une partie substantielle des recherches déposées, sans que l'on puisse en déterminer la représentativité. De plus, on peut considérer que les résumés contiennent le condensé des approches théoriques et méthodologiques mobilisées. En fonctionnant sur le repérage de cooccurrences multiples à l'intérieur de matrices de texte de l'ordre du paragraphe, notre méthode d'analyse doit alors nous permettre de les identifier à partir des résumés.

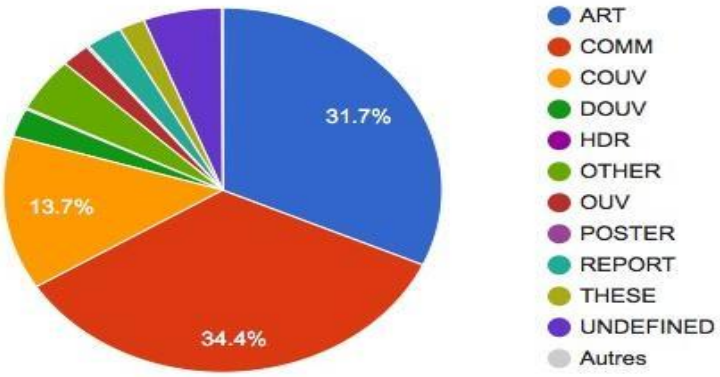
totalité du corpus a une date de publication comprise entre 2002 et 2015 [fig. 1]. Cette situation s'explique aisément par la création d'@rchiveSIC en 2002 (Gallezot et al., 2002) et la date d'extraction (début 2016).

[Figure 1 : répartition des notices par année de publication]



Les notices représentent tous les types de documents mais majoritairement les articles, les actes de colloques et les chapitres d'ouvrages, donc des documents publiés [fig. 2]. Il convient également de noter que plus de 53% des références sont associées à un "full-text", et moins de 47¨% sont donc simplement des références bibliographiques.

[Figure 2 : répartition des notices par type de document]



3 - Traitement et résultats

Notre fouille de texte a deux objectifs : celui d'appréhender les notions liées au temps dans les recherches en SIC et celui de rendre compte du lexique des chercheurs de la discipline dans le temps.

Pour le premier objectif, il s'agit de caractériser le temps tel qu'appréhendé dans les "domaines" ou univers lexicaux (Reinert 1983) identifiés par l'analyse classificatoire. Mais pour révéler ce "temps" il convient auparavant d'avoir un référentiel des notions temporelles. Nous avons choisi de construire ce référentiel à partir de l'appel à communication du XXe Congrès de la SFSIC, dédié à la thématique. Ainsi nous utilisons la même méthode de fouille de texte sur un document rédigé par des chercheurs en SIC qui questionnent le temps. Cela nous permet de prendre la mesure du temps à travers un lexique exogène à HAL. A l'aide de ce dernier nous questionnons un certain lexique lié au temps et sa répartition dans les différents champs de recherche.

Pour le second objectif, nous avons sérié les dépôts en cinq périodes de trois ans pour les rendre quantitativement homogènes sur une durée communément utilisée pour des projets de recherche. Pour chacune des périodes nous observons la place du temps (ie, du lexique du temps issu de l'appel) et le lexique des chercheurs pour distinguer les évolutions.

3.1 - Domaines des SIC

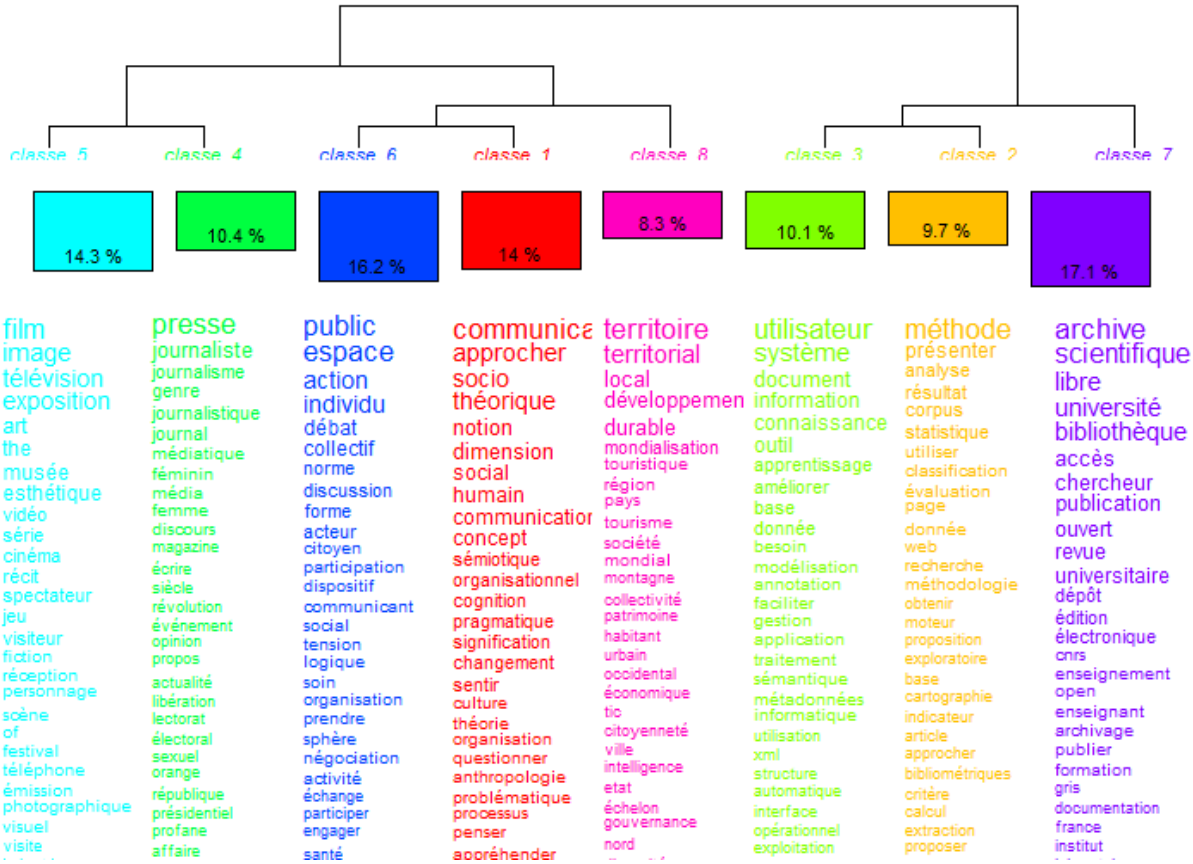
Pour analyser le corpus de 4672 titres et résumés, et caractériser des sous-champ ou domaines, nous utilisons la méthode Reinert (op.cit.) via le logiciel Iramuteq. Une classification hiérarchique descendante (CHD) est réalisée. Elle consiste en un regroupement en classes lexicales des formes cooccurrentes dans des segments de 40 occurrences[7]. Cette première CHD met en évidence l'utilisation non exclusive du français : certains résumés sont fournis en anglais voire dans d'autres langues[8]. La CHD fonctionnant sur la base de la cooccurrence lexicale multiple, ces éléments constituent une classe à part entière et sont isolés du corpus, Iramuteq ne pouvant traiter de corpus multilingue. Le nombre de contributions

7 Pour une description statistique détaillée des procédés, voir Reinert (1983) et Ratinaud et Marchand (2012) pour le fonctionnement spécifique à Iramuteq.

8 A ce sujet, il faut également préciser que la langue de soumission affichée dans les méta-données, renseignée par les déposants, ne correspond pas nécessairement à la langue du résumé, ce qui ne permettait pas d'isoler correctement ces contributions a priori.

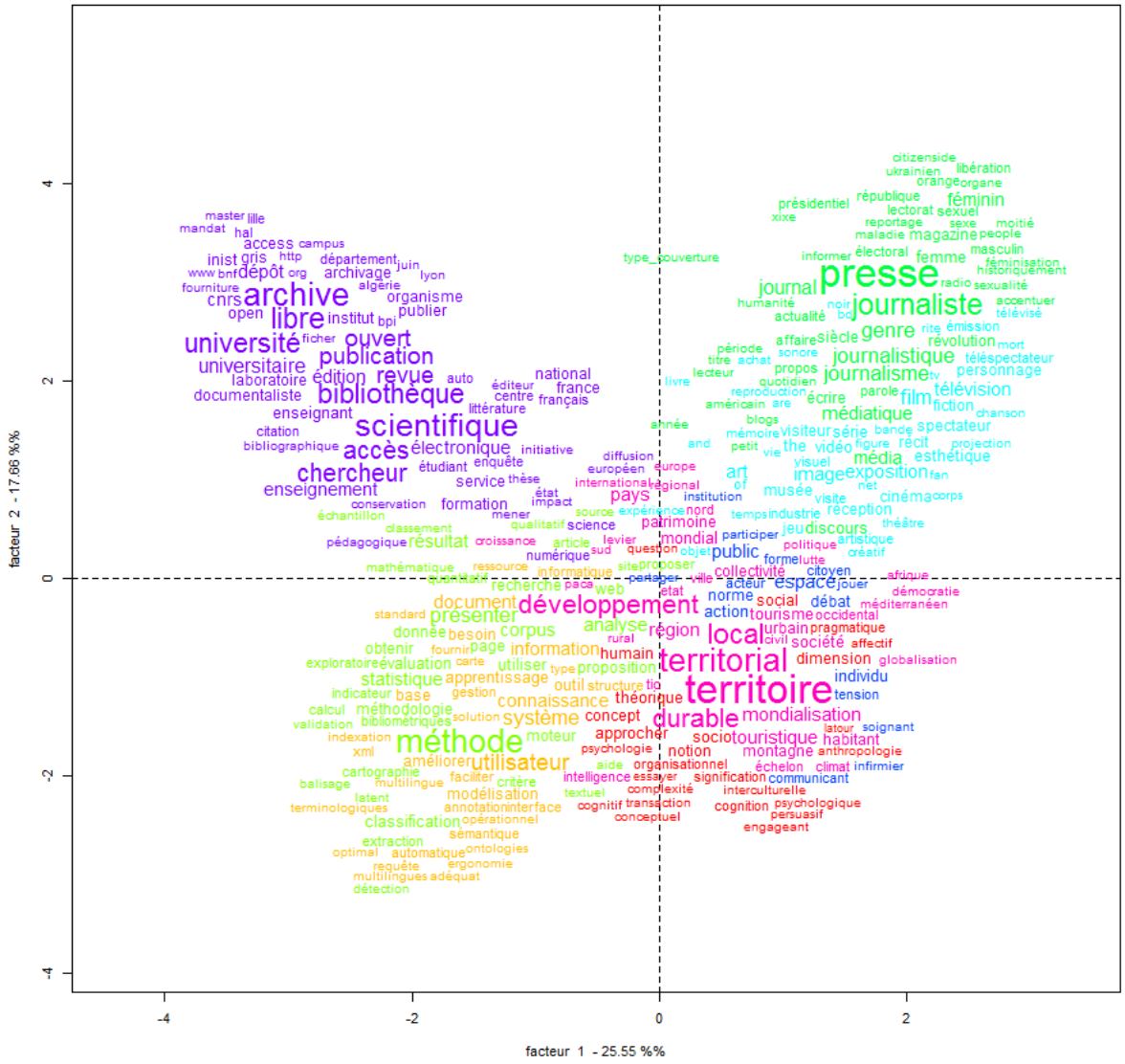
analysées s'élève donc réellement à 4131, sur les 4672 initiales. Une nouvelle CHD est ensuite réalisée sur ce corpus [fig. 3]

[Figure 3 : Dendrogramme de la Classification Hiérarchique Descendante (CHD) sur le corpus en français, avec extrait de leur profil lexical]



Le dendogramme expose 8 classes, représentant les thématiques travaillées par les chercheurs. Une analyse factorielle des correspondances (AFC) [fig. 4] de ces classes permet de "spatialiser" le lexique utilisé par les chercheurs dans les 4131 titres et résumés. Elle met en ainsi évidence les domaines des SIC par le champ lexical utilisé depuis une quinzaine d'années. La CHD et l'AFC mettent en exergue, par la statistique cooccurrentielle, des groupes lexicaux sémiotiquement cohérents. Les termes qui apparaissent en plus gros et plus gras sur l'AFC représentent les termes qui structurent les classes et les discriminent les unes par rapport aux autres.

[Figure 4 : Analyse factorielle des correspondances (AFC)]

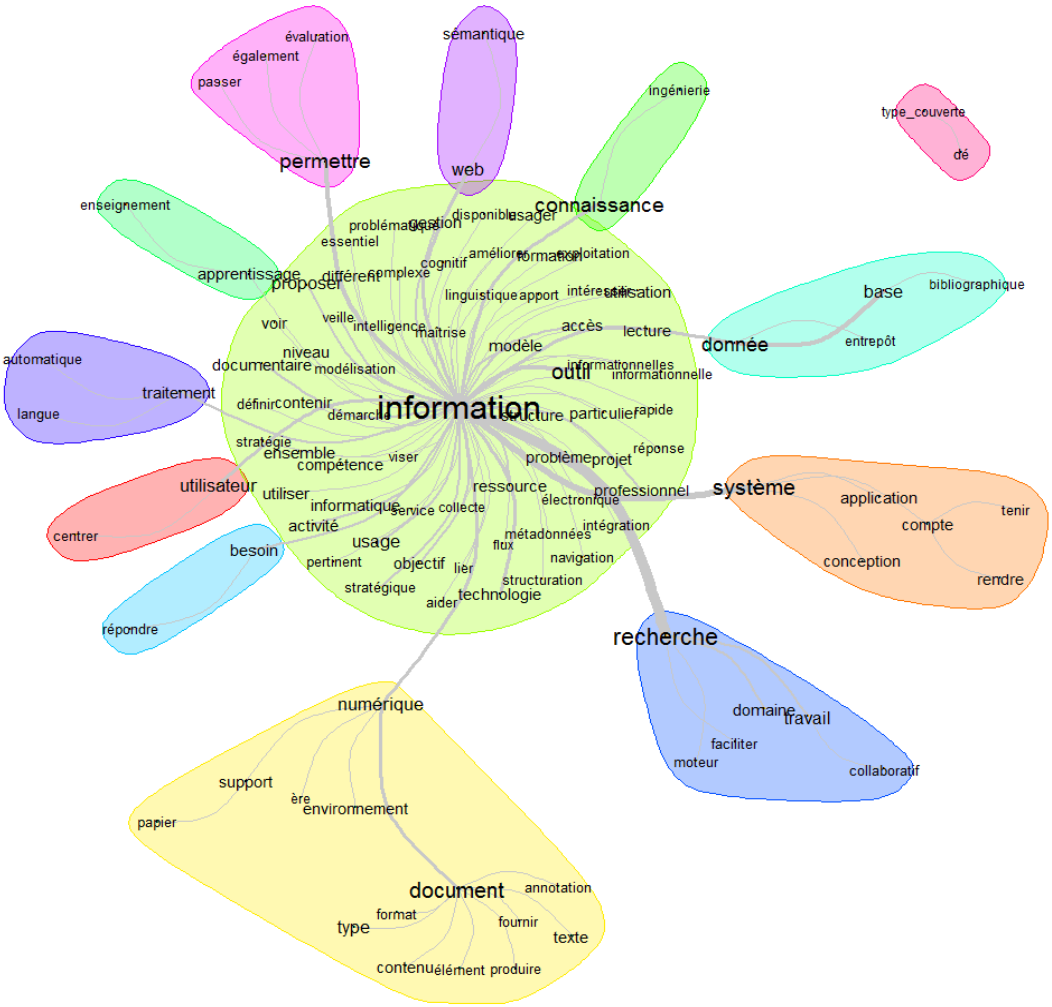


Cette visualisation [fig. 4] permet en premier lieu de montrer les champs de la discipline selon les traces laissées par les chercheurs sur @rchiveSIC. Huit champs sont relativement bien identifiables sur l'AFC, mais une analyse des similitudes à l'intérieur de chaque profil de classe permet de les caractériser de manière plus fine (exemple avec la classe 3 [fig 5]). On peut de la sorte nommer ou étiqueter les classes :

- classe 1 - Les approches théoriques, psycho-sociales ou sémiotiques de la gestion des organisations
- classe 2 - Les méthodologies d'analyse statistique de corpus
- classe 3 - La gestion des systèmes d'information documentaire et le rapport à l'utilisateur

- classe 4 - La presse, le journalisme, les journalistes et/ou leurs discours
- classe 5 - L'image, le film, la télévision et les arts visuels
- classe 6 - L'espace public, l'individu et le collectif, le débat et la participation citoyenne
- classe 7 - Les archives scientifiques, les bibliothèques et l'IST
- classe 8 - Le territoire local, le développement durable, la mondialisation et le tourisme

[Figure 5 : Analyse des similitudes de la classe 3 - Gestion des systèmes d'information documentaire]

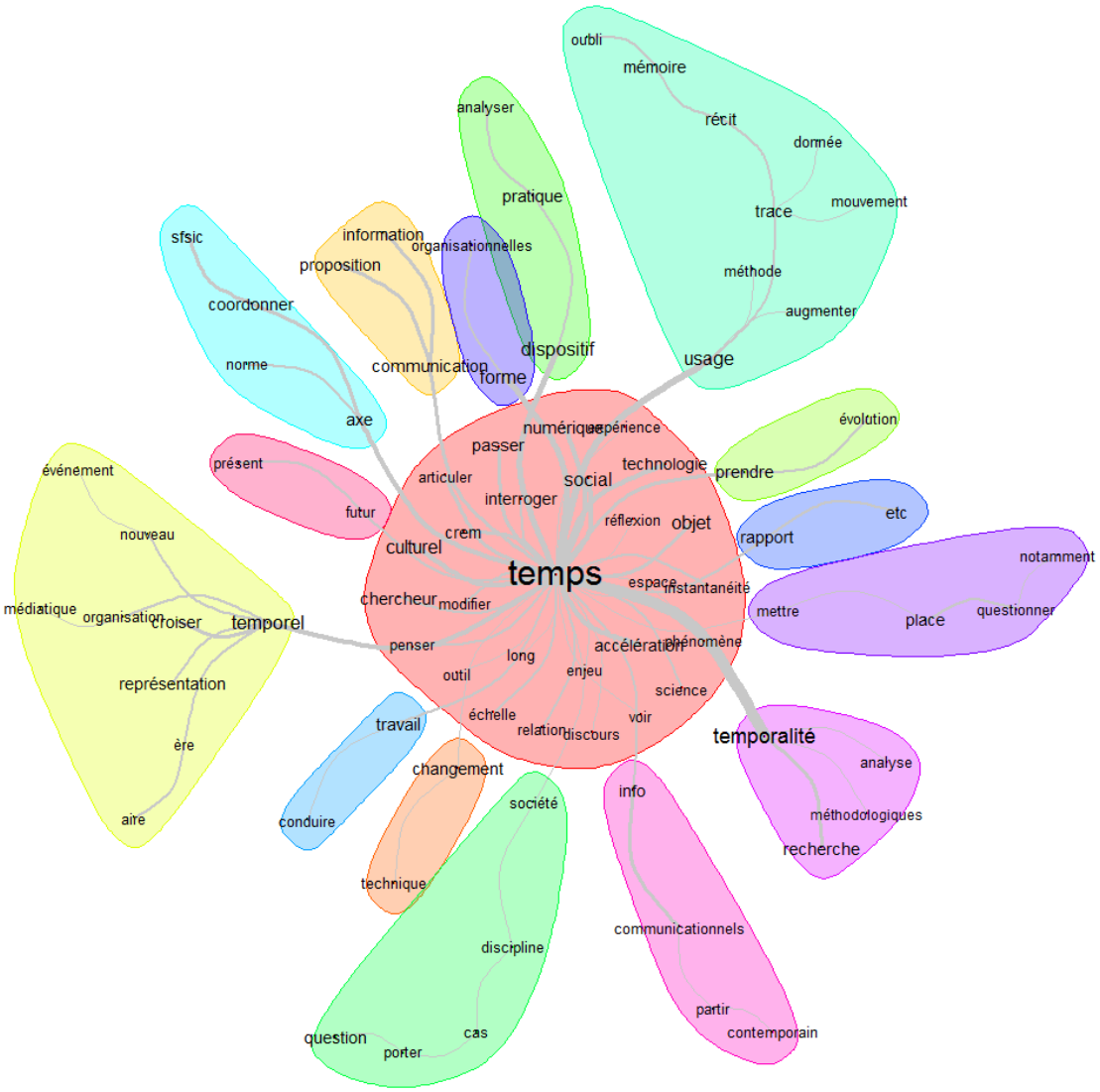


Ces différentes analyses montrent que la question du temps ne caractérise pas une classe en particulier, il faut donc regarder de plus près en allant observer "où" le lexique du temps est le plus utilisé.

3.2 - Référentiel et projection du “temps”

Pour prendre la mesure du temps au sein des SIC, il convient de posséder un lexique de référence exogène au corpus constitué pour le confronter au lexique des différentes classes afin de détecter “par qui” cette question du temps est travaillée. Nous avons choisi pour cela le texte de l’appel à communication, centré sur cette question. Il est synthétisé par l’analyse des similitudes [fig. 6] ci-dessous.

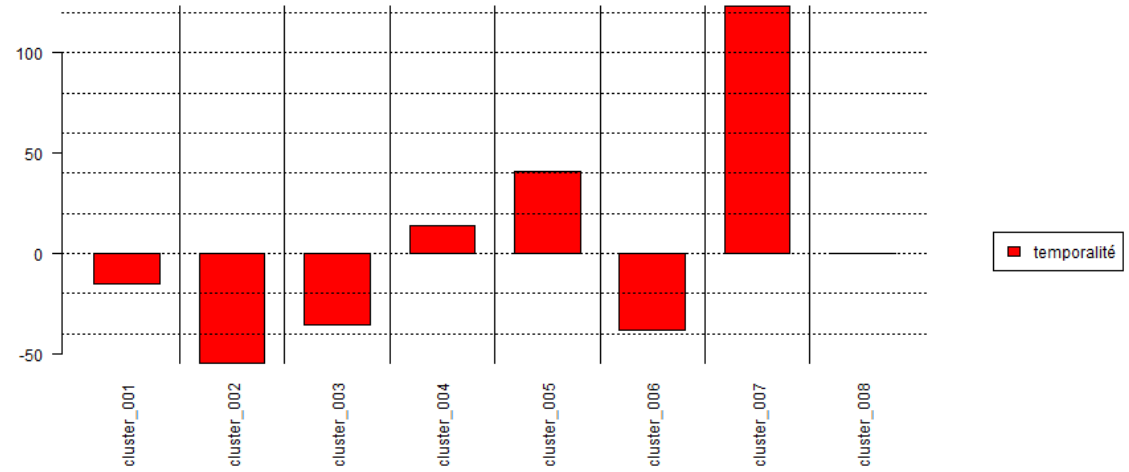
[Figure 6 : Analyse des similitudes de l’appel à communication]



Pour mettre en exergue ce lexique du temps au sein de notre corpus nous construisons un Type Généralisé ou Tgen (Lamalle, Salem 2002) c’est-à-dire un ensemble de formes

lexicales, regroupées manuellement sur la base de critères prédéfinis, ici à partir des "formes actives" (noms, verbes, adjectifs, adverbes et formes non reconnues) de l'appel à communication en retenant les formes lexicales directement connectées à la question du temps[9]. Enfin, nous projetons ce Tgen sur les profils de classes précédemment constituées par la CHD [fig. 7]:

[Figure 7 : Projection du Tgen sur les classes de la CHD]



On observe ainsi que la classe 2 (Méthodologie d'analyse statistique de corpus), la classe 6 (Espace public), la classe 3 (Gestion des systèmes d'information documentaire) et dans une moindre mesure la classe 1 (Approches théoriques de la gestion des organisations) semblent se préoccuper assez peu de la question du temps. A l'inverse, la classe 7 (Archives scientifiques, bibliothèques et IST) semble la traiter plus que les autres, notamment par la présence, logique, des formes *archive, archivage* et *archiver*, mais aussi de *dater*, *fin*, *période* ou encore *récent*. Les classes 5 (Image, film, télévision) et 4 (Presse et journalisme) révèlent elles aussi un léger sur-emploi de ce lexique, respectivement à travers les formes *mémoire*, *histoire*, *temps*, *moment*, *durée*, *rythme* et *année*, *quotidien*, *période*, *histoire*, *agenda*. Quant à la Classe 8 (Territoire et développement durable), elle se caractérise par l'usage moyen de ce lexique de référence (notamment par les formes *échelle*, *anticipation* et *futur*). Dans un

9 Les formes lexicales sélectionnées sont les suivantes: temps, temporalité, temporel, mémoire, futur, accélération, évolution, présent, ère, oubli, long, instantanéité, contemporain, vitesse, urgence, speed, rythme, quotidien, période, mémorielles, mémoriel, moment, histoire, génération, durée, année, éphémérisation, éphémère, uchroniques, synchronisation, spatiotemporelles, slow, récent, ralentissement, ralentir, préservation, présentisme, prédictif, précipiter, échelle, prévision, précipiter, memories, lenteur, immédiateté, hypermodernité, historique, historien, fin, dater, cyclique, court, commémoratif, commencer, avenir, archiver, archive, archivage, anticiper, anticipation, agenda.

second temps, indépendamment de ce Tgen, une lecture attentive des profils lexicaux[10] des différentes classes nous permet d'approcher de manière plus fine les domaines des SIC où la question de la temporalité est saillante, et d'éclairer des notions et appréhensions du "temps" jusque-là passées inaperçues, car non présentes dans l'appel à communication. Ainsi, la classe "journalisme" (classe 4), indépendamment du lexique présent dans l'appel, semble placer la question de certaines temporalités au cœur de ses questionnements, avec la forte présence de formes lexicales désignant des échelles de temps (*siècle*, *année*, *semaine*, *décennie*). La classe 5, centrée sur le film et l'audiovisuel, se caractérise quant à elle par les formes *contemporain*, *époque*, *jour* et *simultanément*. Les recherches ayant trait au territoire local, au développement durable, à la mondialisation et au tourisme (classe 8), comme précédemment esquissé grâce au Tgen, semble se tourner résolument vers le *futur*, avec, outre le lexique déjà mentionné, les formes *futur*, *modernité*, *initier*, *accélérer*, *transition* ou encore *pérennité*. Du côté des classes où le lexique de référence du temps était sous-employé, on notera la présence significative, dans la classe 2 (Méthodologies d'analyse statistique des corpus), des formes *rapide*, *prévoir*, *départ*, *séquence* ou encore *quotidiennement*, clairement liées au temps. Ces derniers résultats pointent d'une part le caractère non exhaustif du Tgen précédemment constitué, et dessinent d'autre part une inscription du temps relativement disparate quantitativement au sein des différents champs de recherche en sciences de l'information et de la communication, et étroitement liée aux objets de recherche privilégiés.

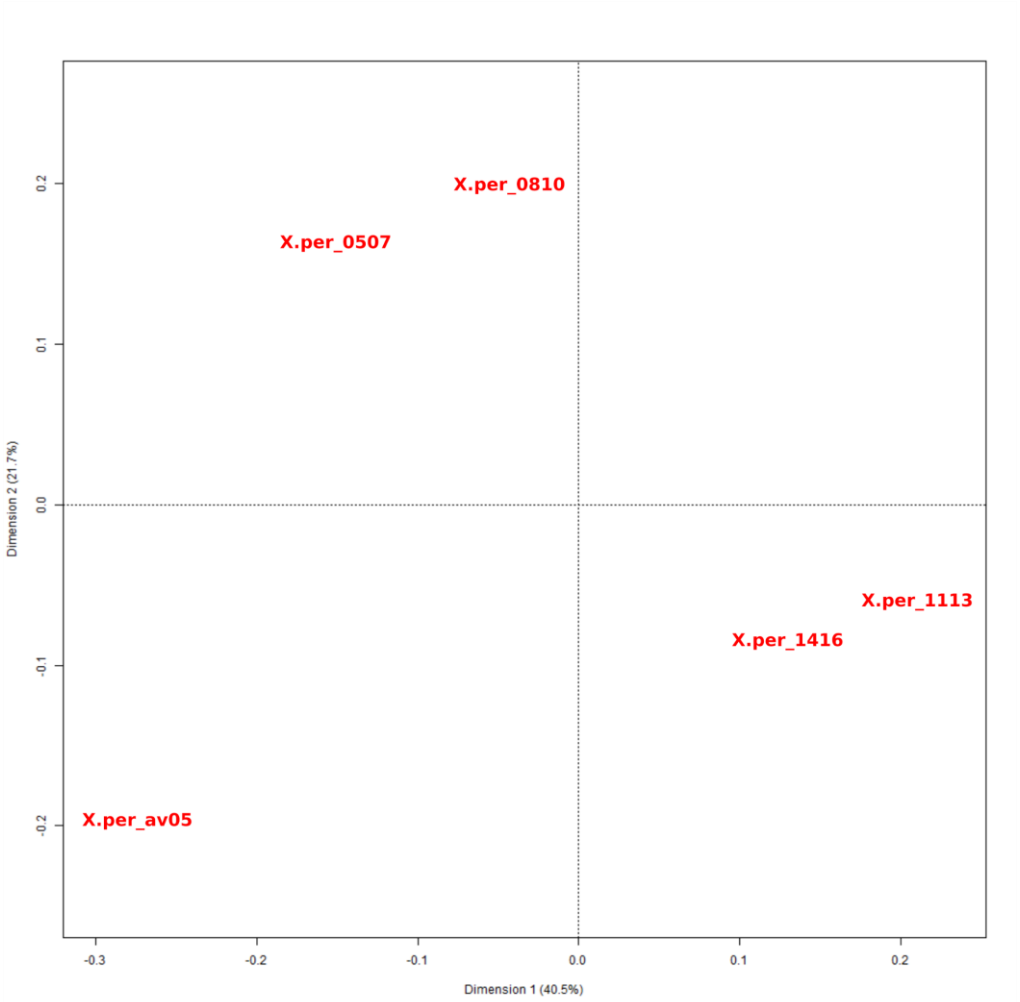
3.3 - Evolution des lexiques

Le corpus est indexé en 5 périodes relativement homogènes en termes de quantité de références : avant 2005 (912 réfs, notée *per_av05*), de 2005-2007 (730 réfs, notée *per_0507*), de 2008 à 2010 (1196 réfs, notée *per_0810*), de 2011 à 2013 (1290 réfs, notée *per_1113*) et 2014 à 2016 (542 réfs, notée *per_1416*).

Le premier élément pouvant nous permettre d'accréditer l'existence d'une évolution diachronique du lexique des dépôts dans HAL est une analyse factorielle des correspondances, effectuée sur les spécificités lexicales des différentes périodes [fig.8].

[Figure 8 : AFC des spécificités lexicales des périodes choisies]

10 Il s'agit des formes lexicales considérées comme spécifiques de chaque classe, dont la liste par ordre décroissant de significativité peut être consultée dans le logiciel Iramuteq. Une petite partie de ces profils est présente dans le dendrogramme (Fig.3).



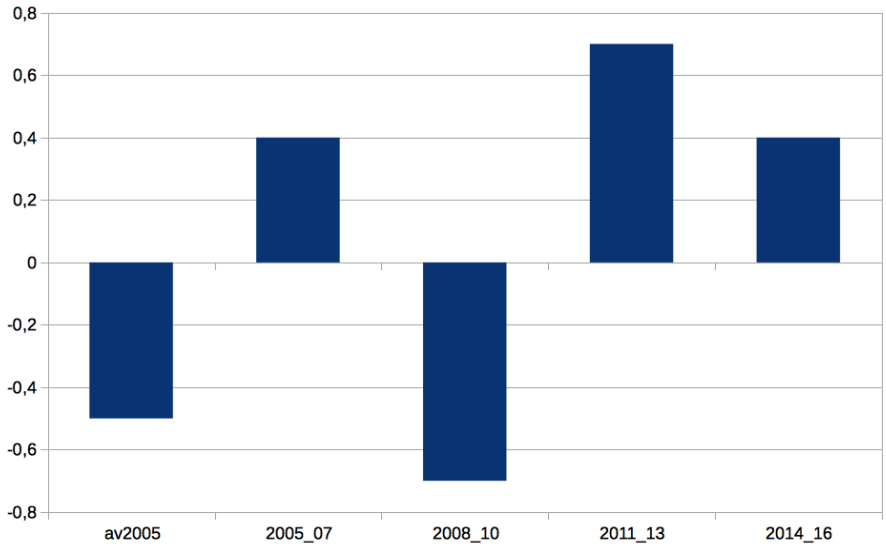
Sur cette AFC, l'axe horizontal, premier axe en termes de significativité statistique[11], voit les différentes périodes s'organiser de manière relativement linéaire. En effet, les périodes les plus anciennes se trouvent à gauche de l'AFC et les périodes récentes à droite, accréditant un glissement progressif du lexique des SIC dans le temps. Seules les périodes 2011-2013 et 2014-2016 semblent à cet égard inversées, mais tout de même situées à droite de l'AFC. Rappelons que l'extraction ayant été effectuée en mars 2016, ladite période est incomplète, ce qui peut expliquer ce léger écart. Néanmoins, la tendance générale exprimée par l'AFC semble bien celle d'une évolution diachronique assez linéaire du lexique. Reste à en déterminer de manière plus précise le contenu et les modalités.

Nous poursuivons donc en observant l'évolution des lexiques : celui du temps représenté par le Tgen issu de l'appel à communication et ceux des chercheurs représentés par les différentes classes. Nous commençons par projeter le Tgen sur les périodes et obtenons ainsi un

11 Exprimée en pourcentage de la variance représentée par l'axe ; ici 40,5%.

graphique [fig. 9] qui permet de visualiser l'évolution du lexique du temps de manière diachronique.

[Figure 9 : Evolution de l'usage du lexique du temps issu de l'appel]



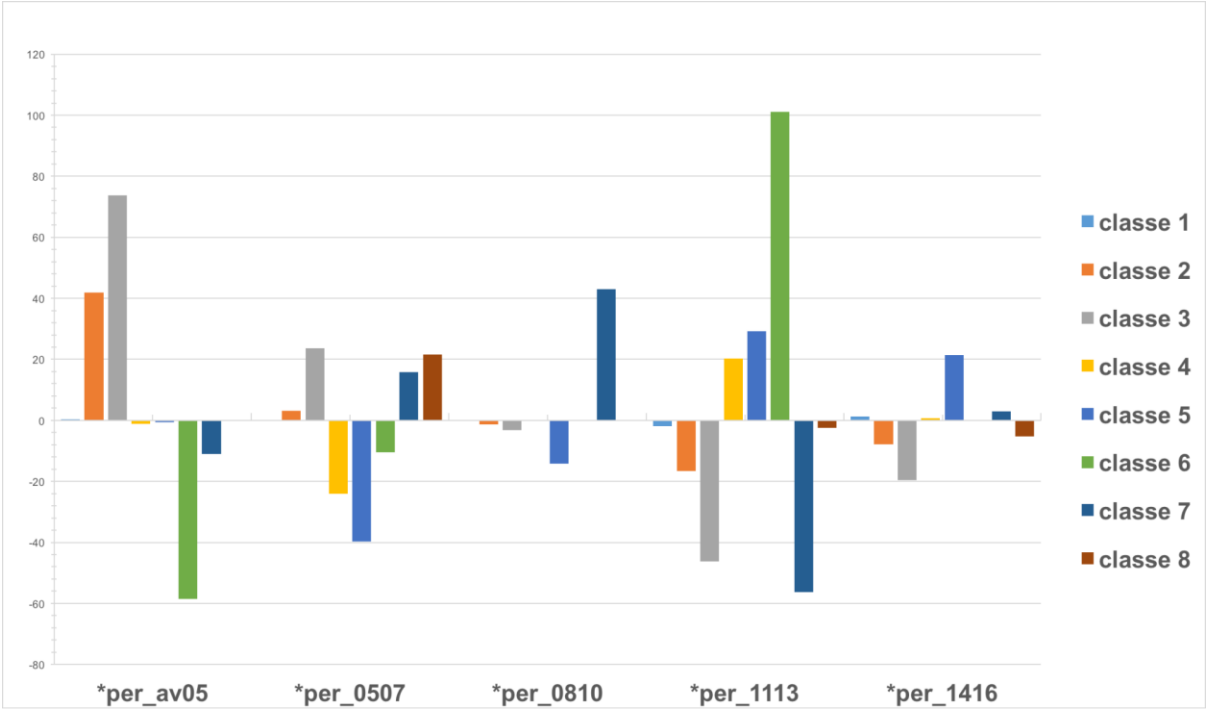
La période antérieure à 2005 et la période allant de 2008 à 2010 sont marquées par un sous-emploi du lexique du temps issu de l'appel. Les périodes où ce lexique du temps a été le plus travaillé sont les années 2005-2007 puis les périodes les plus récentes, depuis 2011 à aujourd'hui. A cette étape il est difficile de dégager une interprétation certaine et univoque de cette distribution. Il convient toutefois d'observer de plus près la période 2008-2010, qui se distingue assez nettement à la fois de la précédente et de la suivante.

Une attention aux types de documents par période révèle une proportion inhabituelle de pré-prints (près de 10%) et « autres documents » (près de 7%) pour la période 2008-2010 quand la période 2011-2013 révèle une quantité plus importante d'articles de revues, d'actes de colloques et chapitres d'ouvrages. Aussi, une première explication pourrait se trouver dans le formalisme de rédaction des textes qui font l'objet d'une publication. Etayer son texte par des références et sur des paradigmes antérieurs ou actuels ou se projeter dans l'avenir sont autant de nécessités qui se matérialisent certainement par l'emploi du lexique du temps observé. Par ailleurs, malgré les erreurs de saisies/indexation dans Hal/@sic, un balayage des thèmes de colloque et des ouvrages et revues laisse entrevoir des thématiques liées au « temps ». On citera par exemple et de manière non exhaustive, pour les revues, « Nouvelles formes de visibilité des individus en entreprise : technologie et temporalité » (*Com&Orga, 2013*), « Les services publics de radio-télévision à l'orée du XXIe siècle » (*Les enjeux de la*

communication, 2013), ou encore « 10 ans de questions de communication » (*Questions de Communication, 2012*).

Le graphique [fig. 10] ci-dessous permet de visualiser l'univers lexical de chaque classe pour les périodes désignées, autrement dit la répartition des domaines des SIC dans le temps. La valeur "0" correspond à une "appartenance" moyenne à la période, les valeurs positives ou négatives à des sur ou sous-représentations du domaine pour la période.

[Figure 10 : Répartition des classes lexicales sur les périodes]



Pour souligner quelques éléments saillants, on peut noter que la période 2011-2013 est fortement marquée par le lexique des classe 4, 5 et 6 (domaines "presse", "média" et "espace public") et la période avant 2005 par les classes 2 et 3 (domaines "Méthodologies" et "système d'information"). On peut s'étonner que des périodes soient marquées de la sorte par un domaine spécifique. Un lissage relatif des domaines, à l'image de la période 2014-2016, est plus intuitif. Une hypothèse interprétative possible serait celle de la non homogénéisation des dépôts dans Hal/@sic. Au gré des constitutions successives des archives institutionnelles, des bilans bibliographiques selon les "vagues" universitaires, les chercheurs issus de tel ou tel laboratoire sont incités ou non à déposer, phénomène pouvant renforcer d'une certaine manière le caractère "thématique" du dépôt.

Il est par ailleurs à présent possible d'avancer une explication du sous-emploi du lexique du temps issu de l'appel pour la période 2008-2010, précédemment observé. Le Tgen est, on l'a vu [fig 7], fortement "présent" dans les classes 7 ("IST") et 5 ("Image et audiovisuel"), indiquant que l'utilisation de ce lexique est accru pour ces classes. Or la période 2008-2010 met en exergue une sur-représentation modérée de la classe 7 uniquement, ainsi qu'une sous-représentation, également modérée, de la classe 5 et un emploi moyen de toutes les autres [fig. 10]. Le lexique du temps présent dans cette période semble donc l'apanage de la seule classe 7, dont le discours, lié aux archives, est assez spécifique. Les autres formes lexicales liées au temps, présentes dans d'autres classes, se trouvent de fait moins représentées que dans les autres périodes, marquées par la présence conjointe de plusieurs classes.

4 – Limites et perspectives

Nous avons caractérisé les domaines des SIC d'un entrepôt d'informations et observé l'intérêt porté par ces domaines à la question du "temps". Nous avons pour cela exploré le champ lexical d'un corpus restreint au titre et résumé en français d'@rchiveSIC que nous avons croisé avec un lexique dédié au "temps". Les résultats obtenus permettent de montrer en contexte l'appréhension du temps par notre discipline. Cette étude, exploratoire, nécessite bien sûr à présent de creuser, d'affiner, de corroborer les résultats obtenus. Ses objectifs initiaux étaient multiples :

- employer, d'abord, des méthodes encore peu ou pas utilisées dans cette problématique de réflexivité disciplinaire, à travers la détection du lexique des chercheurs en SIC et son évolution dans le temps ;
- expérimenter, ensuite, une méthode renouvelable à intervalles réguliers. Il nous semble avoir éprouvé un processus reproductible, y compris pour d'autres thématiques ;
- souligner, également, la nécessité d'alimenter largement ce type d'archives ouvertes et de mieux structurer la discipline avec l'aide de ces plateformes, par un dépôt systématique de la part de notre communauté ;
- proposer, enfin, une méthode généralisable sur la plateforme HAL, idéalement sur le texte intégral des articles, pour y comparer la richesse des questionnements soulevés et des approches méthodologiques mobilisées. La question du TDM (Text and Data Mining) sur un entrepôt ouvert d'information offre beaucoup de perspectives. On entrevoit aussi l'importance de l'exception des TDM pour les plateformes d'éditeurs

dans la loi numérique[12]. Une démarche similaire à la nôtre, mais plus vaste et croisée sur “tous” les entrepôts d’information permettrait de tendre vers l’exhaustivité. On citera par exemple l’outil Gargantext[13] qui permet l’interrogation des corpus d’ISTEX[14], mais également ses propres corpus.

Les limites de ce travail exploratoire sont certaines. La première tient sans conteste à l’impossibilité de constituer un corpus exhaustif des recherches en SIC. Là où certaines disciplines possèdent au moins une base de données bibliographiques centralisatrice (par exemple Pubmed ou PsycInfo) voire des entrepôts d'articles très utilisés (ArXiv ou RePEc), la nôtre n’est pas rendue visible par un dispositif unifié. A cette limite s’ajoutent des difficultés liées à l’auto-archivage, qui complique la normalisation des dépôts (champs de méta-données incomplets, langue de dépôt non respectée, présence ou non du full-text), que cette analyse exploratoire entendait affronter, sans toutefois prétendre pouvoir à ce stade s’affranchir de tout biais.

Ainsi, seuls les chercheurs qui déposent dans Hal / @sic sont représentés. Ceux pour qui la thématique du temps est centrale sont-ils déposants ? Pour avoir une idée de la représentativité du corpus, nous aurions pu observer les laboratoires représentés, mais là encore, il ne peut s'agir que d’un début de réponse. La qualité (la complétude notamment) des données est variable car elle dépend des déposants. Aussi il est assez délicat de prendre en considération un grand nombre de descripteurs sans réduire de manière significative le corpus. L’une des raisons pour lesquelles nous n’avons pas utilisé le texte intégral tient au fait que seul 53% des dépôts en disposent. Si nous réclamions un dépôt massif dans HAL / @sic, nous serions confrontés à la difficulté du traitement multilingue, écarté dans notre étude (4672 > 4131) et qui pourrait se révéler nécessaire sur un corpus plus exhaustif.

Le lexique de référence sur le temps a également été construit par les auteurs à partir d’une logique limitée. Il pourrait sans doute être enrichi par les termes structurant les profils de classe issues de la CHD et relevant de différentes temporalités. Un prochain travail pourrait d'ailleurs consister à constituer sur cette base différents Tgens relevant de différentes temporalités, pour interroger plus finement la question.

Enfin l’énonciation éditoriale (Jeanneret Y., Souchier E., 2005) d’@rchiveSIC et plus globalement de HAL, est bien entendu l’angle mort de notre travail. Mais cette étude

12 Loi numérique : une exception de TDM (presque) "à l'Anglaise" ? http://numeribib.blogspot.fr/2016/07/loi-numerique-une-exception-de-tdm.html

13 http://gargantext.org/

14 http://www.istex.fr/

exploratoire de fouille et d'exploitation des artefacts permet d'autres énonciations, montrant que ce construit peut être recomposé par des outils d'analyse et de visualisation de données capables d'extraire et d'agencer les sédiments pour des néo-documents informés par la statistique textuelle et consciemment subjectivés, premiers jalons d'une méthodologie opératoire pour l'analyse réflexive de notre discipline.

Bibliographie

- Boure R. (dir.), (2002). *Les origines des sciences de l'information et de la communication*: regards croisés Lille : Presses universitaires du Septentrion,.
- Cardy H., Froissart P., (2002). « Les enseignants-chercheurs en Sciences de l'information et de la communication. Portrait statistique ». In : *Les recherches en information et communication et leurs perspectives. Histoire, objet, pouvoir, méthode*. Actes du XIII e Congrès national des sciences de l'information et de la communication Palais du Pharo (Marseille), du 7 au 9 octobre.
- Dumas P., Boutin E., Duvernay D., et Gallezot G., (2006). « Is communication separable from information? » In *First European COnference on communication science*.
- Foucault M., (1969). *L'archéologie du savoir* : Gallimard,.
- Gallezot G., Chartron G., et Noyer J.-M., (2002) « Une archive ouverte des publications en InfoCom ». In , "Colloque SFSIC, "Place et enjeux des revues pour la recherche en InfoCom". Nice .
- Jeanneret Y. et Ollivier B., (2004). *Les sciences de l'information et de la communication*, Hernes N°38, CNRS éditions.
- Jeanneret Y., Souchier E. (2005). « L'énonciation éditoriale dans les écrits d'écran ». Communication et langages.. Vol. 145, n°1, p. 3–15.
- Lamalle C., Salem A., (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *Actes des Sixièmes Journées d'Analyse des Données Textuelles,* St. Malo,.
- Lancien T., Cardy H., Delatte J., Delavaud G., Froissart P., Rodionoff A., Thonon M., Tupper P., (2001). « La recherche en communication en France ». Tendances et carences. MEI - Mediation et information, L'Harmattan.
- Ollivier B., (2000). *Observer la communication: naissance d'une interdiscipline*. : CNRS éd.
- Ratinaud P., Dejean S., (2009). IRaMuTeQ: implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. Presented at the Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009), Toulouse, France.
- Ratinaud P., & Marchand P., (2012). « Application de la méthode ALCESTE à de 'gros' corpus et stabilité des 'mondes lexicaux' : analyse du 'CableGate' avec IRaMuTeQ », Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles, Liège, p.835-844.
- Reinert M., (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. Les cahiers de l'analyse des données.. VIII, (2), 187-198.