

Les SIC à partir du thésaurus Rameau. Représentation ou interprétation ?

par BOUTIN Eric, LIU Pei, GORIA Stéphane, DUMAS Philippe, AMOS David, « boutin@univ-tln.fr »
I3M, LORIA - USTV, Nancy

Les langages d'indexation sont des outils d'aide à la recherche d'information. Ils comportent des concepts (ou vedettes) liés les uns aux autres par des relations sémantiques hiérarchiques ou d'association. Dans cette communication, nous procédons à l'analyse macroscopique des interactions entre concepts du domaine de l'information communication. L'objectif est de considérer un corpus de mots clés français descripteurs du domaine des sciences de l'information et de la communication et de voir de quelle façon ce domaine est représenté dans le langage d'indexation Rameau. La classification du monde de l'information Communication renvoyé par Rameau est elle le reflet de la réalité de la discipline ou une vue sur le monde ? Cette représentation est elle fidèle à la réalité perçue par les chercheurs de la communauté des sciences de l'information et de la communication ? Pour répondre à ces questions, nous proposons de mettre en oeuvre une démarche quantitative de recueil et de traitement automatiques de données relationnelles.

Mots-clés : Langages d'indexation, Rameau, classification, sciences de l'information et de la communication, analyse réseau

In this communication, we proceed to the macroscopic analysis of interactions between concepts in the field of information and communication sciences. The objective is to identify a corpus of french keywords in the field of information and communication sciences and see how this area is represented in the indexing language Rameau. Can we consider that the classification of Information and Communication sciences keywords returned by Rameau is a reflection of the reality or a view of the world ?

Keywords : Indexing languages, Rameau, Classification, information and communication science, social network analysis

De nombreux travaux se sont intéressés à la représentation des Sic en France. Ces travaux ont proposé une représentation macroscopique des Sic (Dumas et al – 2005) ou une représentation particulière d'une de ses composantes (Loneux et al, 2005). Ces travaux ont mobilisé, pour leur partie expérimentale, des corpus scientométriques qui ont fait l'objet de techniques bibliométriques ou qualitatives. Dans ce travail, l'objectif est de proposer une vision macroscopique des Sic en France en étudiant la manière dont notre discipline est appréhendée par le langage d'indexation Rameau. Rameau est un langage documentaire, construit et mis à jour par des bibliothécaires et utilisé pour indexer et donc rechercher les ouvrages de bibliothèques. Les langages d'indexation comportent des concepts (ou vedettes) liés les uns aux autres par des relations sémantiques hiérarchiques ou d'association. Nous entendons par ce travail répondre à plusieurs questions principales :- quelle interprétation du monde des sciences de l'information et de la communication présente le langage d'indexation Rameau ? - comment cette vision des sciences de l'information et de la communication s'est elle affinée au fil des années ?- cette représentation est-elle fidèle à la réalité perçue par les chercheurs de la communauté des sciences de l'information et de la communication ?

Références théoriques de l'étude

La science de la classification a pour objectif de définir les éléments élémentaires univoques de la connaissance et les relations qui existent entre ces éléments de connaissance. Elle est une des approches possibles de l'étude de la relation entre information et connaissance. Notre propos est d'explorer cette approche par l'analyse d'un thésaurus. Mai (2004 b) distingue deux approches théoriques de la classification. - La première est déductive. Les objets sont regroupés dans des

classes sur la base de leurs propriétés observables. Pour Mai (2004a), la structure de la classification doit refléter un ordre préexistant défini a priori. Une classification erronée sera rejetée dès qu'une exception sera identifiée.

Ainsi la classification représente, à terme, la structure vraie de la connaissance. Elle a pour objectif de développer un système qui reflète la réalité (Feinberg-2007). La classification a pour objectif d'encapsuler la variété des perspectives d'un domaine particulier de manière aussi fidèle que possible.- La seconde considère la classification comme une vue sur le monde, une interprétation du monde (Mai -2004a). Cette interprétation du monde dépend de facteurs culturels. Viviane Couzinet (2006) en fournit quelques exemples. La classification russe Bbk (Bibliotечно-bibliograficeskaâ klassifikaciâ) propose plusieurs rubriques consacrées au marxisme léninisme là où les classifications américaines ne consacrent qu'une ligne à ce mot clé. La classification dépend aussi de la perception fine du domaine qu'aura la personne qui procède à la classification.

Un domaine de la connaissance peut être classé selon différentes perspectives épistémologiques. Hjørland (1998) prend l'exemple de la psychologie et précise qu'une classification du domaine de la psychologie doit faire ressortir les approches et les sous disciplines les plus importantes de la psychologie. Cela l'est autant plus que l'on classe des concepts et non des éléments observables comme en sciences naturelles. Il n'y a donc pas une bonne classification qui serait vraie et les autres fausses. Le créateur d'une classification impose une vue particulière de la connaissance. La façon de catégoriser dépend de notre façon de voir le monde (Lakoff-1987). Dans ce travail, nous nous inscrivons dans la logique de la classification comme vue sur le monde et cherchons à analyser la façon dont le vocabulaire des sciences de l'information et de la communication est présent dans le langage d'indexation Rameau.

Méthodologie

Nous allons mobiliser une méthode d'investigation quantitative de type infométrique. La démarche quantitative peut être découpée de façon pédagogique en deux temps correspondant à la constitution du corpus puis aux traitements quantitatifs qui lui sont associés. *La collecte des données de l'analyse quantitative* : Le langage d'indexation Rameau est constitué de concepts reliés entre eux par des relations sémantiques. Lorsqu'on interroge l'interface web de Rameau[1], on peut connaître pour un concept donné son (ou ses) termes génériques, son (ou ses) termes spécifiques et son (ou ses) termes associés. Le langage Rameau est un langage d'indexation vivant qui est modifié à partir des propositions formulées par le réseau de ses utilisateurs. Lorsqu'on souhaite effectuer une proposition, celle-ci est transmise au correspondant de l'établissement inscrit au Fichier National des Propositions. Les grandes bibliothèques universitaires et les grands Services Commun de Documentation ont leur correspondant. Cette proposition est alors traitée par le gestionnaire Rameau du domaine considéré. Les sciences de l'information et de la documentation sont un domaine de Rameau : elles ont à ce titre un gestionnaire de domaine[2] « La communication » se retrouve dans différents domaines sans apparaître jamais de façon explicite. Les propositions de modifications prennent la forme d'ajouts, de modifications ou de suppressions de concepts. Elles sont enregistrées dans une base de données interrogeable appelée *fichier national des propositions* Rameau[3]. Dans le domaine des sciences de l'information et de la documentation, 96 modifications ont été proposées par 22 établissements différents. Le tableau 1 présente les 11 établissements qui réalisent 85% des demandes de modification.

Etablissement ayant suggéré la modification	Nombre de modifications demandées
CFCB Caen	49
BNF D4	4
ENSSIB	4
HE Spaak	4
BDP Dordogne	3
BM St-Claude	3
BNF BFL 1	3
BNF MANUSCRITS (division orientale)	3
BNFD4	3
Indexpresse	3

Tableau 1 : 11 correspondants réalisent 85% des propositions de modifications dans le domaine des sciences de l'information et de la documentation

Il est ainsi possible de connaître, pour un concept donné, quand et par qui ce terme a été proposé à l'ajout, à la modification ou à la suppression. Ces modifications, si elles sont validées par le gestionnaire de domaine, sont intégrées dans le langage d'indexation Rameau. Il est ainsi possible de connaître, pour chaque terme du thésaurus, quand, par qui et pourquoi il a été introduit. Ces informations sont publiées chaque semestre dans le « journal des créations et des modifications » et constituent une matière première précieuse. Elles sont consultables au format Pdf. pour les révisions de Rameau postérieures à Mars 1996[4]. Rameau présente l'avantage de garder à tous les niveaux une traçabilité des différentes modifications qu'il a connues au fil des versions. Ces données permettent alors de conduire une analyse dynamique de type longitudinale.

Notre travail a consisté à collecter de façon fédérée, pour un ensemble de termes des Sic les informations disponibles sous ces différentes interfaces. Ces informations ont été présentées autour de deux tables. Une table des interactions qui se présente sous forme d'une liste de concepts- et une table qui présente les relations hiérarchiques entre ces concepts. Pour déterminer la liste des concepts de la discipline des Sic en France, nous aurions pu faire l'intersection entre une liste reconnue de mots clés de notre discipline et les concepts contenus dans Rameau. Notre corpus de départ a été construit d'une autre manière. Il a consisté à considérer au départ trois mots clés (information, communication et sciences de l'information et de la communication) et à retenir les termes dans l'environnement de ces trois mots clés avec une profondeur de trois. Le problème posé par cette technique d'identification des concepts est qu'elle négligera peut être certains mots clés du domaine des Sic qui ne sont pas dans le voisinage proche de ces trois termes. Ce processus d'exploration en profondeur est décrit figure1.

Information
Communication
SIC

Figure 1 : Le processus d'exploration en profondeur

Nous tenons à préciser que cette exploration avec une profondeur de trois s'effectue en prenant les descendants directs d'un concept (son concept générique et le concept générique du générique, idem en aval) mais également ses ascendants ou descendants indirects. Pour prendre une analogie généalogique, nous allons considérer les parents, grands parents, arrières grands parents, enfants, petits enfants, arrières petits enfants mais aussi les oncles, grands oncles, cousins et petits cousins. Ceci conduit à enrichir considérablement la taille de notre corpus, Rameau se caractérisant comme un « petit monde »

Le traitement des données quantitatives

Les données ainsi constituées font l'objet d'un traitement en utilisant l'analyse des réseaux sociaux (Wasserman et Faust, 1994). Des représentations cartographiques sont proposées qui permettent de représenter le nuage de points des mots clés descriptifs des Sic au sens de Rameau. Ces représentations cartographiques sont ensuite filtrées par date afin de reconstituer la dynamique de création de la discipline dans Rameau. Les indicateurs de l'analyse réseau sont également mobilisés pour décrire la position de chaque concept au sein du réseau et son évolution au fil du temps. Nous avons choisi de décrire la position de chaque sommet en utilisant des indicateurs de centralité. Les analyses en terme de réseaux sociaux, pour leur part, ont mis en évidence cinq formes de centralité : centralité de degré, centralité de proximité, centralité d'intermédiation, centralité au sens de l'information, centralité de flot. Elles sont présentées par Degenne et Forsé (1994).

Nous avons privilégié deux de ces indicateurs. Ce choix se justifie d'une part par l'interprétation intuitive à laquelle ces deux indicateurs peuvent donner lieu ainsi qu'à leur facilité d'être calculé (en interne ou en recourant à des logiciels d'analyse réseau). La centralité de degré considère qu'un sommet *i* est plus central qu'un sommet *j* si le sommet *i* a plus de sommets qui lui sont adjacents que le sommet *j*. Le concept de centralité de degré est facile d'obtention. Il peut être décomposé en deux indicateurs correspondant à Indegree et Outdegree. Indegree correspond au nombre de liens entrants sur un sommet et outdegree au nombre de liens sortants. Cet indicateur donne une vision très locale de la centralité d'un sommet puisqu'il ne prend en compte que les sommets adjacents.

Toutefois, dans la pratique, on observe que cet indicateur conduit à un classement très proche de celui d'autres indicateurs beaucoup plus sophistiqués. Nous avons retenu aussi la notion de centralité au sens de l'intermédiarité. Un sommet *i* sera d'autant plus central qu'il appartiendra à un grand nombre de chemins géodésiques. Le chemin géodésique entre deux sommets *j* et *k* est le plus court chemin entre ces deux sommets. Lorsqu'un sommet se situe sur plusieurs géodésiques, cela signifie qu'il est le point de passage obligé entre de nombreux sommets du réseau. Cette situation peut s'apprécier également en terme de contrôle. Si *i* appartient au chemin géodésique entre *j* et *k*, cela signifie que *i* contrôle l'interaction entre *j* et *k*. Pour calculer cet indicateur, il faut considérer l'ensemble des géodésiques correspondants aux paires de sommets (*i*, *j*) du réseau. Pour obtenir l'indicateur de centralité au sens de l'intermédiarité associé au sommet *i*, il faut compter le nombre de fois où le sommet *i* apparaît dans les géodésiques. Ce second indicateur est calculé par le logiciel Netdraw.

Nous avons découpé notre corpus en périodes de cinq ans. Pour chaque période, nous avons calculé la centralité des sommets. La centralité d'un sommet évolue au fil du temps par l'ajout ou la suppression d'autres sommets qui se trouvent connectés ou déconnectés de lui. La centralité d'un sommet ne s'apprécie pas par son ancienneté mais par son positionnement relatif aux autres. Il n'y a pas de prime à l'ancienneté, un concept très récent peut apparaître directement plus central qu'un concept plus ancien. Ces évolutions peuvent être traduites de façon graphiques et rendre compte des termes de la discipline qui sont les plus centraux, ceux qui évoluent vers une position de plus en plus centrale. Les résultats obtenus sont-ils fidèles à la réalité perçue par les chercheurs de la discipline ?

Résultats de l'expérimentation

L'étape de collecte des données nous a permis d'identifier 1063 relations entre les 904 concepts qui sont situés à un niveau de profondeur de 3 maximum des trois concepts retenus (information, communication, Sic). La représentation de ces 1063 interactions ne donne pas une figure très lisible. Nous avons fait le choix, pour simplifier le graphe d'éliminer les sommets qui n'avaient de relations qu'avec un seul autre sommet. Les sommets résiduels (195) ont été représentés sur un réseau visualisé

Annexe 1. Nous avons représenté sur ce réseau les points d'articulation du graphe qui établissent des relations entre des univers qui seraient sinon disjoints. Les points d'articulation du graphe sont représentés en bleu. En découvrant ce graphe on peut faire plusieurs observations :

- Il y a une sur-représentation des concepts plutôt associés au monde de l'information et des bibliothèques. Peut être ce biais est-il dû au fait que les indexeurs des concepts sont issus du monde de la documentation.
- Dans le répertoire Rameau, la « vedette » Sciences de l'information est employée pour sciences de l'information et de la communication et vice versa. On est donc face à une représentation du monde dans laquelle Rameau ignore la conjonction de deux champs qui se sont mis à dialoguer depuis 30 ans.
- Ces deux mondes sont représentés schématiquement au Nord et au Sud du graphe et le nombre d'interactions entre eux est faible. Cette séparation entre information et communication est confirmée lorsqu'on projette ces résultats sur un espace réduit en utilisant la technique de Multidimensionnal scalling. L'axe des ordonnées est le plus discriminant. Les 10 concepts correspondant aux valeurs polaires de ce graphe sont présentés tableau 2.

Concepts au nord de la MDS	Concepts au Sud de la MDS
Bibliothèques et multimédias	Culture
Multimédias interactifs	Enseignes
Algorithmes	Signes et symboles

Fichiers (informatique) — Organisation	Arts graphiques
Systèmes de gestion de données techniques	Bibliographie matérielle
Types abstraits de données (informatique)	Conversation
Bibliothéconomie — Logiciels	Droit — Édition
Bibliothèques et Internet	Livres — Industrie et commerce
Échange électronique d'information	Logotypes
Logiciels	Marques typographiques

Tableau2 : Représentation des 10 concepts opposés dans la MDS La figure 2 permet également d'illustrer une relation entre les vedettes matières information et communication qui pose question. Communication apparaît comme le terme générique de sciences de l'information. Terme spécifique : Sciences de l'information

Terme générique :Communication

Bibliologie

Documentation

Recherche de l'information

.....

Figure 2 : résultat du zoom sur information Après avoir conduit cette première analyse relationnelle, nous avons cherché à représenter les interactions entre les concepts par tranche d'année. Nous avons réparti notre corpus en 6 périodes de cinq années chacune. La première période commence en 1980 et considère tous les concepts de 1980 à 1984. La période 6 est tronquée aux concepts intégrés depuis 2005. Nous avons considéré tableau 3 6 périodes de 5 ans chacune.

Périodes	dates	Nombre de concepts introduits
Période 1	80-84	101
Période 2	84-89	65
Période 3	89-94	31
Période 4	94-99	34
Période 5	99-04	10
Période 6	04-08	3

Tableau3 : découpage temporel de notre corpus

L'activité de création de nouveaux concepts n'a pas été linéaire avec le temps comme l'indique la dernière colonne du tableau 2. La figure 3 traduit le même phénomène et permet de visualiser la diminution de l'introduction de nouveaux concepts. Cette évolution dénote une maturité des sciences de l'information et de la communication, une stabilisation du vocabulaire.

Figure 3 : évolution de l'introduction de nouveaux concepts au cours du temps

A partir de ces 6 sous corpus, on peut construire 6 réseaux que l'on peut superposer mais les résultats ne sont pas très exploitables. Nous avons pour chacune des ces périodes calculé les deux indicateurs de centralité que nous avons retenus. Les concepts sont alors présentés suivant l'évolution de leur niveau de centralité au fil du temps. On a classé tableau 4 les 10 mots clés qui ont les centralités d'intermédiarité les plus fortes : on observe une surreprésentation de la dimension documentaire.

	Création	Dernière mise à jour
Systèmes d'information	1981/09/01	2004/08/11
Informatique documentaire	1984/08/21	2002/01/22
Réseaux d'information	1981/09/01	2001/02/12
Information électronique	1996/07/03	2006/07/28
Bibliothéconomie	1982/02/09	1995/11/28
Sources d'information électroniques	1999/05/03	1999/11/13
Bibliologie	1984/06/22	1995/09/22
science de l'information	1984/08/21	1996/02/16
Services de documentation	1982/02/10	2004/08/10
Technologie de l'information	1988/05/27	2003/10/01

Tableau4 : 10 mots clés ayant l'intermédiarité la plus forte

On a représenté figure 4 l'évolution de quelques concepts lorsqu'ils sont classés par indicateur d'intermédiarité. Pour raisonner sur des bases identiques, nous avons centré et réduit les valeurs de centralité obtenues. On représente dans cette figure les vedettes matière selon que leur valeur de centralité d'intermédiarité centrée réduite a progressé ou régressé entre 1984 et 2008. A la droite de la figure, on note des mots clés qui ont régressé en terme de centralité entre les deux périodes.

Figure 4 : vedettes matières classées par évolution de leur centralité centrée réduite en 1984 et 2008

Conclusion

A l'issue de ce travail, nous pouvons apporter quelques éléments de réponses à la question que nous avons posée au départ. Le langage documentaire Rameau offre-t-il un reflet de la réalité ou une interprétation du monde ? Les résultats que nous avons pu mettre en évidence nous invitent à une réponse claire sur cette question. L'analyse a mis en évidence d'une part une confusion entre sciences de l'information et Sic, d'autre part un rattachement ambigu des Sciences de l'information à la communication et enfin une survalorisation des vedettes matières correspondant à des termes associés au domaine des sciences de l'information en général et aux métiers de la documentation en particulier. Le langage d'indexation Rameau est donc loin de favoriser une lecture des différentes sensibilités existant dans le domaine des sciences de l'information et de la communication en France. Nous avons convenu de soumettre les résultats à une analyse qualitative auprès d'experts de la communauté des Sic en France. Cette analyse qualitative aurait été le moyen de confronter la vision du monde offerte par Rameau à des chercheurs de la communauté des Sic. Toutefois, le manque de finesse des résultats dans le domaine des sciences de la communication n'aurait pas permis une critique constructive. Nous avons renoncé pour cette raison à conduire cette analyse qualitative. Le langage d'indexation Rameau aurait pu être utilisé comme outil de veille des concepts émergents dans notre communauté. Les résultats de ce point de vue sont également décevants : les nouvelles entrées dans le langage documentaire Rameau sont de plus en plus rares. Ce travail renforce la nécessité de créer un thésaurus de la discipline qui représente ses différentes sensibilités. D'autres thésaurus francophones auraient pu être explorés pour cette étude et notamment *Termscience*[5]. Toutefois ce thésaurus apporterait un éclairage sur la composante informationniste de notre communauté mais en aucun cas sur sa composante communication.

Bibliographie

Courbières C., 2000. *De la mode et des discours au regard de l'indexation documentaire*. Thèse de doctorat nouveau régime, sciences de l'information et de la communication, université de Toulouse-le Mirail.

Couzinet V., (2006) « Les connaissances au regard des sciences de l'information et de la communication : sens et sujets dans l'inter-discipline » *Journée ISKO-France dans le cadre de la Semaine de la connaissance* 26/06/2006

Degenne A. Forsé M., (1994), *Les réseaux sociaux*, Editions Armand Colin.

Dumas P., Boutin E., Duvernay D., Gallezot G., (2005) "Is Communication Separable from Information", *Proceedings of First European Communication Conference*, Amsterdam, 2005, [en ligne] Disponible sur : http://archivesic.ccsd.cnrs.fr/docs/00/06/27/07/PDF/sic_00001670.pdf

Feinberg, M., (2007). Beyond retrieval : A proposal to expand the design space of classification. *Proceedings of the North American Symposium on Knowledge Organization*. Vol. 1. Disponible sur : <http://dlist.sir.arizona.edu/1892>

Hjørland B., (1998). The classification of psychology : a case study in the classification of a knowledge field. *Knowledge Organization*, 25 (4) : 162–201.

Lakoff, G. (1987). *Women, fire, and dangerous things : What categories reveal about themind*. Chicago : University of Chicago Press.

Loneux C., Bourdin S., Bouillon JL, (2005) Building the field of organisational communication in France : concepts, methods, institutions, First european communication conference, Amsterdam, 2005Mai

Jens Erik, (2004-a), classification of the web : challenges and inquiries, *Knowledge Organization*, 31, N° 2, 2004, P 92-97Mai

Jens Erik, (2004-b), classification in context : relativity, reality and representation, *Knowledge Organization*, 31, N° 1, 2004, P 39-48

Wasserman S., Faust K. (1994). *Social Network Analysis : Methods and Applications* : Cambridge University Press

Annexe 1 : représentation des points d'articulation du graphe

[1] http://catalogue.bnf.fr/jsp/recherche_autorites_rameau.jsp;jsessionid=0000Va70O_kqJICEeMhftgxMu5t:-1?nouvelleRecherche=O&host=catalogue

[2] La liste des gestionnaires par domaine peut être consulté à l'adresse <http://rameau.bnf.fr/informations/domaines.htm>

[3] <http://rameau.bnf.fr/utilisation/introduction.htm>

[4] <http://rameau.bnf.fr/utilisation/journal.htm>

[5] <http://www.termssciences.fr/-/Index/Rechercher/Rapide/>