

Rapport d'étape

Jean-Philippe Cointet

2 mai 2008

Sujet:

Dynamiques sociales et informationnelles dans les réseaux épistémiques:
morphogenèse et diffusion

sous la direction de:

Paul Bourgine (CREA, ISC)

Pierre-Benoît Joly (TSV)

1 Titre de la thèse et rappel des objectifs

Titre: *Dynamiques sociales et informationnelles dans les réseaux épistémiques : morphogenèse et diffusion*

Directeurs de thèse:

- Paul Bourguine, CREA: Centre de Recherche en Epistémologie Appliquée (Ecole Polytechnique - CNRS), ISCIPLIF (Institut des Systèmes Complexes Paris Ile-de-France)
- Pierre-Benoît Joly, TSV: Transformations Sociales et politiques liées au Vivant (INRA)

Cette thèse s'attache à décrire, analyser et modéliser les réseaux épistémiques définis comme un système d'acteurs en interaction au sein d'un réseau social qui manipulent, échangent et produisent de l'information. L'hypothèse fondamentale adoptée est que la structure et la dynamique de ces systèmes socio-sémantiques est co-déterminée par les interactions entre agents et par la distribution des connaissances. A titre d'exemple on peut citer l'organisation des communautés scientifiques dont la structure (traduite dans les réseaux de collaboration par exemple) est en partie guidée par les affinités thématiques respectives des chercheurs (ex: comportement d'homophilie/hétérophilie lors de la création de nouvelles collaborations) qui peuvent en retour influencer l'évolution des champs thématiques.

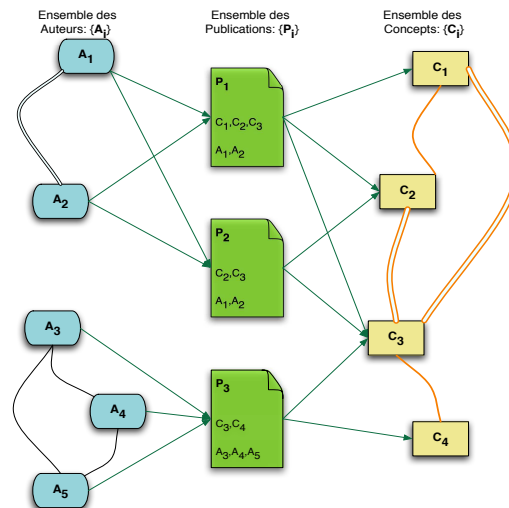


Figure 1: Un exemple de réseau épistémique : l'activité scientifique. Des acteurs interagissent au sein d'un réseau de collaboration scientifique, tandis que la distribution des connaissances est ici formalisée par un réseau de co-présence de concepts dans les publications.

La formalisation des systèmes sociaux sous forme de réseaux permet de les appréhender simultanément à plusieurs niveaux. En effet, les réseaux se définissent par l'ensemble des interactions interindividuelles qui correspondent aux "dynamiques microscopiques" du système, mais la mise à jour de leurs structures - au sens des motifs non triviaux qui les caractérisent - est également décisive pour comprendre "leur évolution macroscopique". Les deux niveaux sont à nouveau étroitement liés, les structures macroscopiques émergeant des comportements individuels (morphogenèse), et ces dernières contraignant les dynamiques microscopiques. Dans le cas de l'activité scientifique, les chercheurs produisent sans cesse de nouvelles collaborations (bas niveau), dont émergent des motifs structurels tels que les communautés scientifiques dotées d'une dynamique de haut niveau autonome, qui peuvent elles-même avoir un effet contraignant sur les prochaines collaborations. Cette boucle de rétroaction bas-niveau/haut-niveau, émergence/immersion peut encore être enrichie par l'introduction de la dimension sémantique, indispensable pour appréhender la complexité des dynamiques à l'oeuvre dans les réseaux épistémiques. Le schéma 2 résume ces nombreux aspects entre-mêlés en couplant les dimensions sociales et sémantiques à tous leurs niveaux.

prochaines sections illustrent chacun un projet de mise en rapport de ces méthodes avec des objets et des questions propres aux sciences sociales: (i) un travail de cartographie des dynamiques scientifiques dans le contexte du renouvellement paradigmatique introduit par les approches réseaux - complexité en biologie qui s'insère dans le projet COBINA (Connaissances biologiques et Normes d'Action Publique), (ii) un projet qui porte sur l'évolution des réseaux de collaboration qu'entretiennent les experts internationaux dépêchés au sein de comités du Codex Alimentarius.

2 Cartographie des dynamiques scientifiques, généalogie des approches réseaux et complexité en biologie.

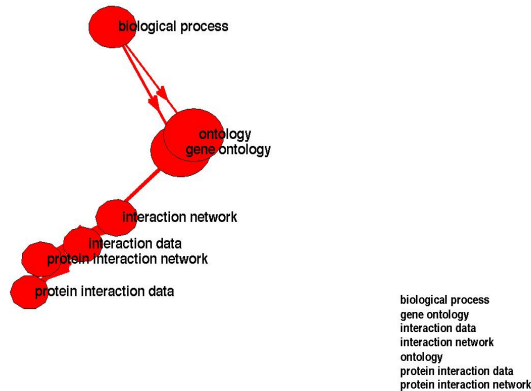
La représentation des dynamiques scientifiques constitue un premier exemple d'application des méthodes développées au cours de la thèse. L'objectif est de construire une représentation multi-niveau des dynamiques conceptuelles à partir de statistiques basiques de cooccurrences de termes extraites à partir de bases de données scientifiques. Plus spécifiquement les méthodes relèvent à la fois de la fouille de données, de la scientométrie, et de l'analyse de réseaux.

Les publications scientifiques constituent une trace de l'activité de recherche de communautés scientifiques. Les chercheurs communiquent notamment leurs résultats sous la forme de publications scientifiques. C'est une voie de communication stigmergique en ce qu'elle est ouverte, non orientée, et pérenne (la numérisation des archives amplifiant encore ces caractéristiques par rapport aux archives papier). Les publications cristallisent un état de la connaissance à un moment donné ; chaque nouvel article provoque de nouvelles associations entre des termes parfois familiers parfois étrangers, et est susceptible de modifier le paysage conceptuel d'un champ de recherche. Cette information est néanmoins bruitée, partielle et se trouve distribuée sur les millions d'articles publiés annuellement, par des communautés parfois parfaitement déconnectées. L'ambition est de reconstruire une structure de la connaissance hiérarchisée à partir de l'ensemble de ces informations distribuées.

La méthodologie est la suivante. Etant donné un corpus de textes datés (ici des publications scientifiques), et un corpus de termes librement défini, on extrait le nombre de cooccurrences n_{ij}^t observées entre chaque paire de termes dans les abstracts des articles publiés une année donnée. Cette information de base sert à mesurer une proximité sémantique entre les termes. La mesure employée pour calculer cette proximité sémantique se différencie des distances employées classiquement en scientométrie en ce qu'elle est asymétrique (les mesures traditionnelles considèrent généralement des métriques du type $d(i, j) = d(j, i) = \frac{n_{ij}}{n_i n_j}$, dans notre cas, on introduit un paramètre de focalisation α qui permet de rendre compte de l'hétérogénéité de la distribution des termes: $d_\alpha(i, j) = (\frac{n_{ij}}{n_i})^\alpha (\frac{n_{ij}}{n_j})^{1/\alpha}$). On en déduit deux types de voisinage: le voisinage en spécificité qui réunit les termes qui spécifient le terme cible, et un voisinage en généralité qui permet d'extraire les contextes dans lesquels le terme cible peut s'inscrire.

Une matrice de proximité entre termes est ensuite construite. Elle définit un réseau orienté et pondéré sur l'ensemble des termes, le poids d'une arête w_{ij} de ce graphe étant égale à la proximité sémantique entre i et j . Ce réseau sémantique est ensuite traité avec des algorithmes de détection de communautés (au sens d'ensembles de noeuds localement denses) avec recouvrement qui permettent de réduire le réseau à un niveau mésoscopique composé d'un ensemble de "champs paradigmatiques" et de leurs articulations. L'utilisation d'un algorithme de détection de communautés recouvrantes permet de traiter de façon automatique la polysémie des termes. Si un terme peut prendre différentes connotations en fonction de son contexte, il apparaîtra autant de fois qu'il existe de contextes à même de le spécifier. La procédure peut être répétée sur le niveau mésoscopique pour reconstruire de façon émergente le niveau supérieur... On itère l'opération à chaque pas de temps pour observer l'évolution de la structuration des champs scientifiques au fil de leur diversification, fusion, scission, etc.

Un "champ paradigmatique" se caractérise donc comme un ensemble de termes fortement interdépendants à une période donnée. L'asymétrie de la mesure de proximité permet de plonger cet ensemble de termes dans un référentiel local hiérarchisé. Chaque terme est doté de coordonnées dans un espace à deux dimensions: en ordonnée, on calcule la moyenne des distances entrantes de ce terme à l'ensemble de ses voisins, et en abscisse,



Field number 139 : 2004-2007

Figure 3: Champ paradigmatique des réseaux d'interaction génétique et protéinique. On voit que le terme "biological process", bien que faisant partie à part entière du champ a une position encore marginale dans sa structure générale.

la moyenne des distances sortantes à destination de l'ensemble des termes du champ. Cette projection permet d'apprécier la structuration interne d'un champ. Ainsi les termes bien inscrits dans le champ global se positionnent sur la bissectrice principale, en bas à gauche pour les termes les plus précis, en haut à droite pour les plus généraux, tandis que les termes plus annexes ou en cours d'incubation par le champ ont des positions plus périphériques (voir exemple figure 3).

Cette méthode de reconstruction des dynamiques scientifiques a été appliquée dans le cadre du projet ANR "COBINA" (Connaissances biologiques et Normes d'Action Publique) en collaboration avec des historiens des sciences (Christophe Bonneuil et Jean-Paul Gaudillère) à un corpus de publications en biologie dans le contexte du renouvellement paradigmatique ouvert par les méthodes d'analyse de la biologie systémique et la prise en compte croissante de l'importance de l'épigénétique et des grands réseaux d'interaction.

Un ensemble de plus de 800 termes a ainsi été défini et cartographié (cf figure 4) sur les cinquante dernières années afin d'illustrer les mutations profondes subies par la biologie sur cette période. Les méthodes de reconstruction des dynamiques scientifiques ont ainsi pu être validées par des experts du domaine. Mais l'objectif principal est d'accompagner le travail de reconstruction historique en fournissant une représentation visuelle de la structuration des sous-domaines de la biologie à une époque donnée, et de leurs dynamiques (cf figure 5). Les cartes ainsi produites peuvent servir de support à l'exploration d'hypothèses expliquant les mutations épistémologiques contemporaines.

La carte figure 4 fait par exemple clairement apparaître un noyau central en rouge foncé (taux de croissance du champ supérieur à 100% par rapport à la période antérieure) qui correspond à des outils et techniques d'analyse. Cette position centrale dans la période la plus récente illustre le rôle prépondérant des outils et instruments matériels dans l'analyse de réseau tendant à confirmer que la percée récente de l'approche réseau repose fondamentalement sur des bases matérielles (expérimentale et bio-informatique) issues de la biologie à haut débit. Au centre, les champs 164-169-177 correspondent ainsi à l'analyse de profils d'expression des gènes (puces à ADN et ARN=microarray), tandis que plus bas les champs 135-139-158 sont liés aux données d'interactions protéines et annotations de bases génomiques. Les aspects écologie/évolution (en haut à gauche) ne sont encore connectés que de façon superficielle au coeur instrumental dur, même si on peut observer l'émergence de communautés absentes des périodes précédentes (118: evolvability, 113: gene regulatory network/evolution, 69: evolution/scale-free, voire 85: variance/simulation...) qui font le lien entre la communauté instrumentale centrale et les communautés plus classiques de l'évolution et du développement. On retrouve également des approches nouvelles dans les outils théoriques mis en oeuvre (en bas à gauche, champs 35, 143 et 144 réseaux de neurones artificiels, SVM, algo-

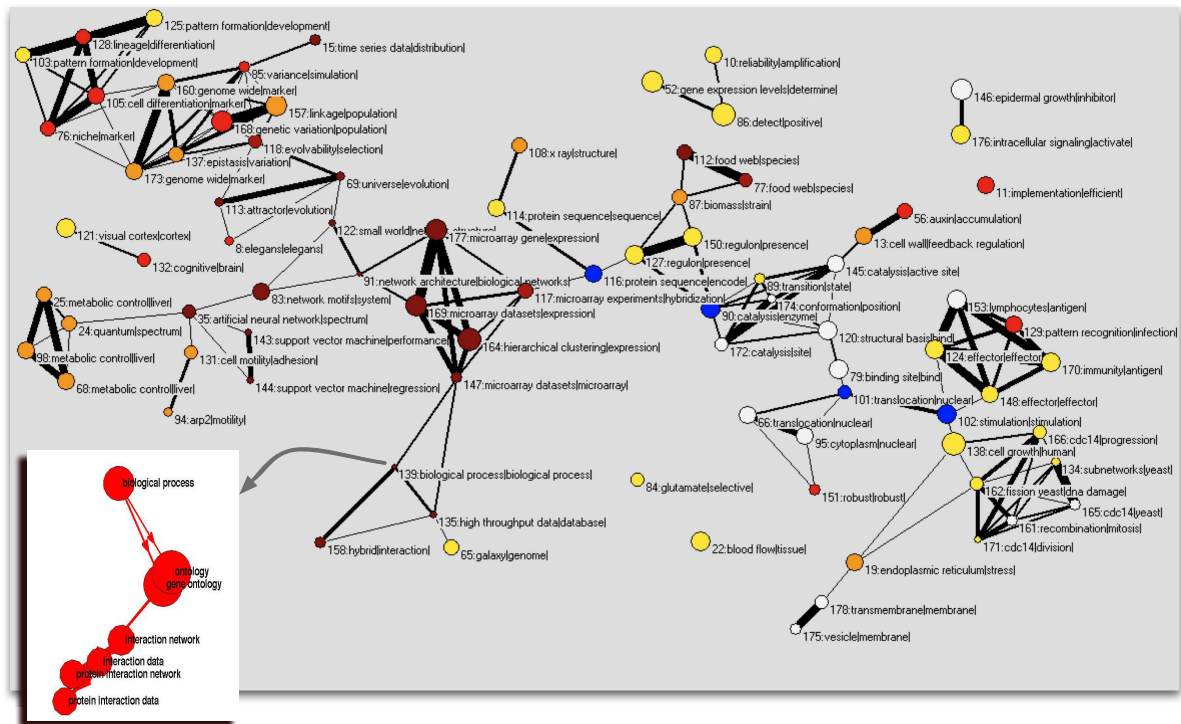


Figure 4: paysage conceptuel de la période 2004-2007 autour du thème biologie et réseau. Chaque noeud du réseau représente un champ paradigmatique illustré par un exemple dans l'encart en bas à gauche. La taille d'un champ est proportionnelle à son activité tandis que sa couleur reflète la croissance de son taux d'activité (bleu: décroissance de plus de 50%, blanc, champ stable, jaune, croissance de 50%, rouge, croissance supérieure à 100%.

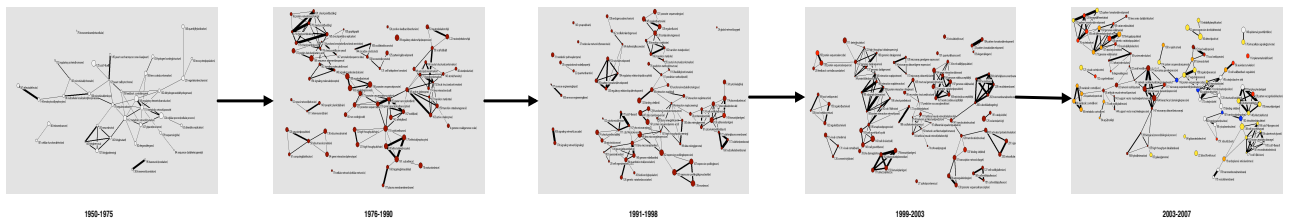


Figure 5: Evolution de la structuration des champs sur cinq périodes.

rithme génétique...). Les approches écosystémiques sont également très actives, et voient leur composition évoluer (champs 77, 112: présence de réseaux trophiques).

Une dernière approche, encore exploratoire, complète l'analyse de cartes telle qu'elle a été esquissée ci-dessus: la description en termes de dynamiques mésoscopiques. La méthode de cartographie permet de reconstruire une structure multi-niveau de l'organisation des termes pour une période donnée. La juxtaposition de cartes produites à différentes périodes permet de comparer la structuration générale des champs mais elle n'offre pas la possibilité d'étudier les dynamiques à plus petite échelle. L'objectif est maintenant de décrire la dynamique discrète des champs paradigmatiques reconstruits. Ce travail permet de tracer la généalogie des champs en détectant les événements de fusion/scission de champs, naissance, mort, croissance, décroissance. Dans l'exemple illustré figure 6, on a tracé la phylogénie des champs depuis 1963 en coloriant en rouge l'ensemble des champs contenant le terme réseau. La polysémie du terme apparaît ici clairement. Cette représentation permet de retracer finement l'histoire des champs en identifiant les influences croisées entre sous-domaines.

Les outils développés ont donc permis d'objectiver et d'accompagner les hypothèses sur une transition du tout génétique à une vision moins déterministe des mécanismes biologiques. Les questions propres à la représentation

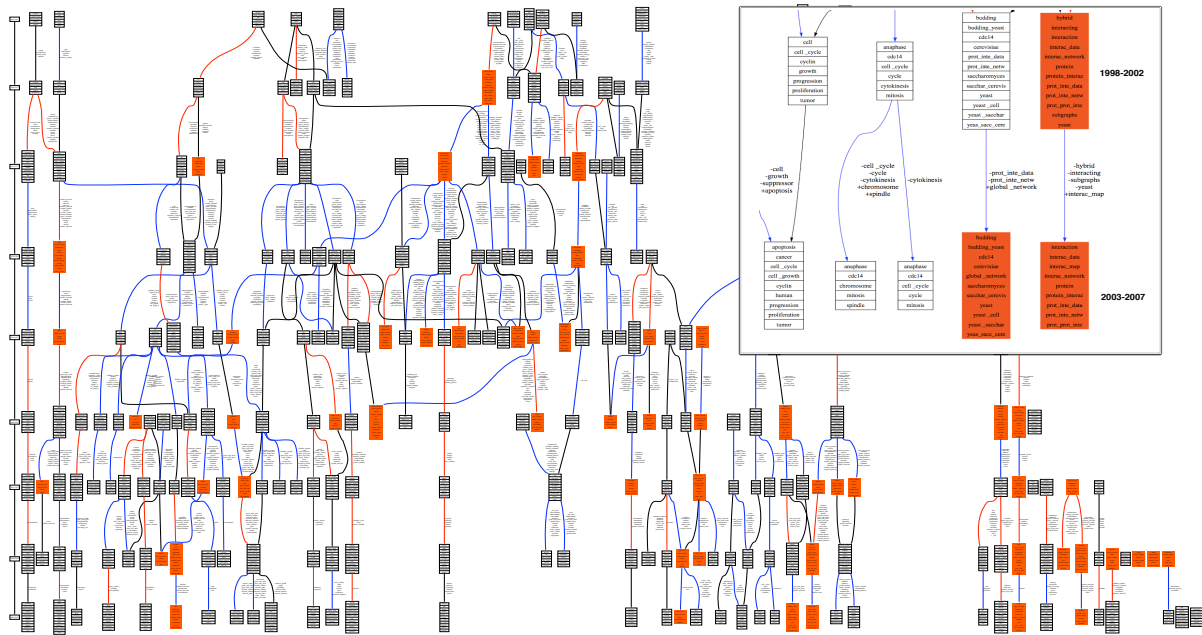


Figure 6: Phylogénie des champs paradigmatiques de 1963 à 2007. L'ensemble des champs contenant le terme réseau sont coloriés en rouge, chaque ligne représente l'ensemble des champs et leur composition sur une période de quatre ans. Un lien bleu signifie qu'il y a eu globalement décroissance de la communauté, un lien rouge, croissance (enrichissement du nombre de termes). Les labels le long des liens signalent les termes acquis et perdus d'une période à la suivante. Les événements possibles sont: naissance (absence de mère), mort (absence de fille), croissance (une mère, lien rouge), décroissance (une mère, lien bleu), scission (une mère plusieurs filles), fusion (une fille, deux mères). L'insert en haut à droite est un zoom sur les deux dernières périodes.

de connaissance multi-échelle et en dynamique sont ici cruciales autant à cause de la complexité des données à mettre en forme que par la nécessité d'intégrer les spécialistes dans la boucle de modélisation. Cette dernière contrainte a été prise en compte en mettant à disposition des experts un outil de navigation spécifique qui prend la forme d'un site web, permettant de naviguer à travers les différentes périodes et à travers les différents niveaux selon la résolution souhaitée. Enfin, des possibilités d'annotation permettent de partager les interprétations et remarques (cf. figure 7).

3 Production d'expertise scientifique internationale au sein de comités du Codex Alimentarius

Depuis 1963, le Codex Alimentarius, une organisation intergouvernementale de l'OMS et de la FAO, propose un cadre international à la sécurité sanitaire des aliments par l'élaboration de textes normatifs. Ces textes d'application volontaire (standards, normes, lignes directrices et guides de bonnes pratiques entre autres) sont produits au sein de comités constitués de délégations des pays membres où les discussions donnent une place privilégiée aux données scientifiques et techniques ainsi qu'à la recherche de consensus. Ces comités sont donc des lieux stratégiques où sont discutées et décidées les procédures d'évaluation des risques.

Anotine Debure s'intéresse à ces processus de prises de décisions et de construction de standards qui sont traversés d'enjeux politiques, économiques, sociaux et scientifiques. C'est dans ce contexte que nous cherchons à décrire la composition de ces comités indépendants saisis par les comités CODEX.

La méthodologie que nous avons adoptée est la suivante. Nous travaillons sur deux comités indépendants: le JECFA (Joint Experts Committee on Food Additives) et le JEMRA (Joint Experts Meeting on Microbiological Risk

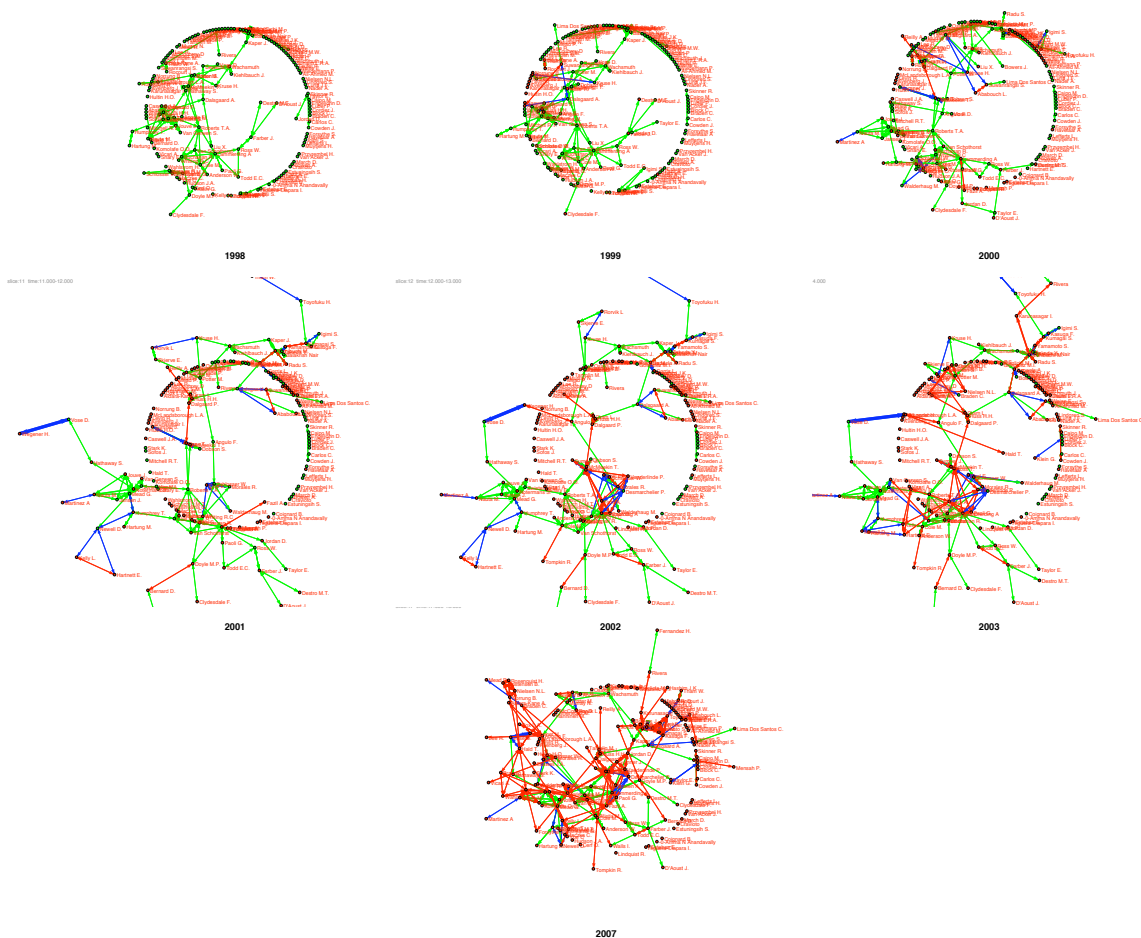


Figure 9: Réseaux de collaboration des membres du JEMRA de 1998 à 2003 puis en 2007, les noeuds rouges représentent les experts ayant déjà participé à une consultation, les noeuds verts les membres à venir, une vision dynamique permet de constater quelles collaborations préexistaient à la participation aux consultations (liens verts ou bleus), et quelles collaborations sont postérieures (liens rouges).

De la même façon, ces analyses permettent d’infirmer ou de confirmer des hypothèses sur les conséquences observées de la participation des chercheurs à ces consultations sur leur activité scientifique “traditionnelle”. On admet généralement que la participation à ces comités donne accès à un certain nombre de ressources (aussi bien en termes d’identification de points chauds qu’en termes d’enrichissement du capital social) qui sont ensuite à même d’être valorisées dans d’autres arènes.

La recherche de collègues invisibles qui préexisteraient à la création des comités ne se limitent pas à la recherche de communautés de collaborateurs, mais plus largement à l’identification de communautés épistémiques définies comme des ensembles d’individus partageant les mêmes domaines d’expertise scientifique. Une analyse des publications sous forme de réseaux socio-sémantiques et des outils de reconstruction de ces communautés épistémiques sous forme de treillis de gallois doivent permettre de mieux comprendre la dynamique de construction de ces comités dans leur double dimension scientifique et sociale.

4 Publications

Publications dans des revues internationales à comité de lecture

- J-P. Cointet, C. Roth (2007) "How Realistic Should Knowledge Diffusion Models Be?". *Journal of Artificial Societies and Social Simulation*, 10(3):5, 2007, 20 p.
- D. Chavalarias, J-P. Cointet. (2008), "Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study" *Scientometrics*, 75(1), 2008, 14 p.
- J-P. Cointet., D. Chavalarias (2008), "Multi-level science mapping with asymmetrical paradigmatic proximity" *Network and heterogeneous media*, 3(2), June 2008, 11 p.

Publications dans des conférences internationales à comité de lecture

- J-P. Cointet, E. Faure, C. Roth "Intertemporal topic correlations in online media", in *Proceedings of the 1st ICWSM International Conference on Weblogs & Social Media*, Boulder, Col., E.-U., mars 2007, 4 p..
- J-P. Cointet, C. Roth, "Information diffusion in realistic networks", in *Actes d'AlgoTel 9e rencontres francophones sur les aspects ALGOarithmiques des TELécommunications*, Ile d'Oléron, mai 2007, 4 p. Élu "Best Student Paper"
- D. Chavalarias, J-P. Cointet, "Science mapping with asymmetric co-occurrence analysis: methodology and case study on the complex systems community", *ECCS 2007, Dresden, Oct. 2007*
- L. Tabourier, J-P. Cointet, C. Roth (2008), "Cycles in hypergraph-based networks: signal or noise, artefacts or processes ? ", in *Actes d'AlgoTel 9e rencontres francophones sur les aspects ALGOarithmiques des TELécommunications, Ile d'Oléron, mai 2008, 4 p.*

Workshops et Conférences à processus de reviewing à base d'abstracts

- C. Roth, J-P. Cointet, "Knowledge diffusion models", *Sunbelt XXVII International Social Network Conference, Corfu, Greece, May 1-6, 2007*
- J-P. Cointet, C. Taramasco, C. Roth, "Socially-mediated concept diffusion in a scientific community", *Sunbelt XXVII International Social Network Conference, Corfu, Greece, May 1-6, 2007*
- Sallantin J. et al "A Logical Framework to Annotate Documents in a Virtual Agora", *The square of opposition, Montreux, Switzerland, June 1-3, 2007*
- D. Chavalarias, J-P. Cointet, "Dynamical Science Mapping", *Sunbelt XXVII International Social Network Conference, Corfu, Greece, May 1-6, 2007*
- J-P. Cointet "Opinion dynamics in the french political blogosphere" *Scaling in Biological and Social Networks, SFI - Santa Fe, NM, USA - July 2007*
- J-P. Cointet, C. Roth "Evolution des structures relationnelles et des contenus au sein de la blogosphère politique française", *Nouvelles approches, nouvelles techniques en analyses de réseaux sociaux, CLERSE, Lille, Mar. 2008*
- J-P. Cointet "Modèles de diffusion de connaissance, quel réalisme?", *atelier TICOOP, Nice, Fév. 2008*

5 Participation à des enseignements et encadrement de stagiaires

- "Multiple dynamical networks", *V Summer School Complex Systems Institute: ISCV*, Valparaíso, Chile, January 2008
- mars-sept 2006: Stage long (6 mois) de master 2: Carla Taramasco (master de sciences cognitive),
- mars-sept 2007: Stage long (6 mois) de master 2: Pierre Chèvremont (master de sciences cognitive),
- mai-juillet 2007: 3 Stages courts (3 mois) de l'ENPC: Hugo Lebrun, Richard Norton et Jean-Charles Cizel,
- mai-juillet 2008: 2 Stages courts (3 mois) de l'ENPC: Flavien Moreau et Pierre-Olivier Ricault,
- janv-avril 2008: Stage court (3 mois) de master 2: Mathieu Galvez (master AIV)
- mai-août 2008: Stage court (3 mois) de master 1: Adrien Friggeri (ENS Lyon)