# Reconstruction multi-échelle des dynamiques scientifiques*

Jean-Philippe Cointet [†,‡]
Jean-Philippe.Cointet@polytechnique.edu

10 août 2008

**Abstract**

L'activité scientifique est essentiellement dynamique et processuelle. Nous proposons des méthodes de reconstruction automatisées de ces dynamiques grâce à des mesures de proximité entre termes construites depuis des bases de données scientifiques. L'objectif est de cartographier les sciences à partir d'informations très partielles distribuées sur des millions d'articles. L'originalité de ce travail est d'introduire une reconstruction multi-échelle des sciences qui rende compte de leur structure interne et qui intègre pleinement la dimension temporelle dans l'analyse. Un cas d'étude concernant la biologie des réseaux est présenté et brièvement commenté.

## 1 Introduction

L'activité scientifique est composée de la somme des processus complexes [10] issus de réseaux d'interactions hétérogènes mêlant chercheurs, ingénieurs, objets d'expérimentation, outils, journaux, institutions, etc...[12]. La publication scientifique, dont la validité est attestée à l'issue d'un processus d'évaluation par les pairs est généralement considérée comme un des produits principaux de ces interactions multiples. Les centres d'intérêt de chaque communauté scientifique sont ainsi cristallisés au sein des publications. On peut considérer que chaque article modifie de faÁon même infinitésimale l'état de la connaissance à un moment donné en provoquant de nouvelles associations entre des concepts parfois familiers, parfois étrangers.

Les publications scientifiques agissent également comme un des modes de communications principaux entre chercheurs. C'est une voie de communication stigmergique [9] car ouverte (même si cette propriété est plus caractéristique de la période récente et est très liée à la numérisation des archives), indirecte, et pérenne. Elle permet la coordination à longue distance des chercheurs en leur offrant un état de l'art sans cesse réactualisé de l'activité

*Ce travail a été développé dans le cadre du projet COBINA ("Connaissances biologiques et Normes d'Action Publique") financé par le programme OGM de l'ANR.

†TSV (INRA, France)
‡CREA (Ecole Polytechnique, France)

scientifique. Il nous paraît donc naturel de nous attacher au même matériau d'origine pour reconstruire les dynamiques scientifiques.

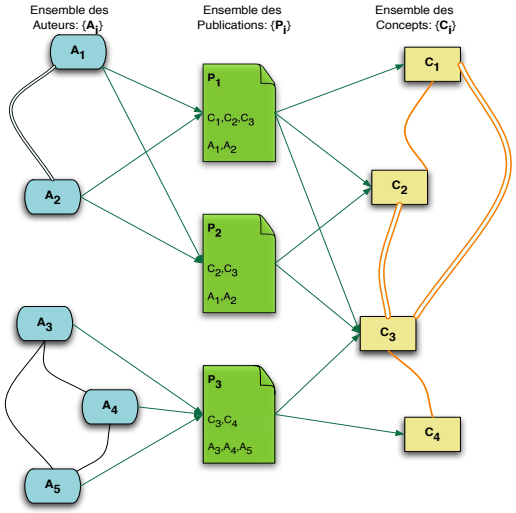


Figure 1: Les chercheurs interagissent au sein d'un réseau de collaboration scientifique, tandis que la distribution des connaissances est ici formalisée par un réseau de co-apparition de concepts dans les publications.

La figure 1 représente le processus de base de l'activité scientifique. Des chercheurs $\{A_i\}$ produisent des publications $\{P_i\}$ qui mettent en relation des concepts $\{C_i\}$. En suivant l'hypothèse de Kuhn selon laquelle un "paradigme est constitué de ce que partagent les membres d'une communauté scientifique, et inversement, une communauté scientifique est peuplée d'individus partageant le même paradigme"[11] [1], nous postulons qu'il existe un isomorphisme très fort entre la structure des communautés scientifiques, et la structure conceptuelle associée à une distribution préférentielle d'usages de termes observée dans les publications. Nous considérerons également que les statistiques basiques sur les cooccurrences de termes sont un bon marqueur de ces structures.

Nous nous intéressons donc à la dynamique des communautés scientifiques, en nous concentrant sur les "champs paradigmatiques" entendus comme l'ensembles des termes (outils, objets, concepts) employés préférentiellement dans ces communautés. On tracera les glissements thématiques des communautés scientifiques en repérant des associations inédites entre termes, ou la disparition de certaines.

La reconstruction des dynamiques des communautés scientifiques est un enjeu à la fois méthodologique et théorique. Appréhender l'évolution des sciences à partir de données réelles présente un intérêt particulier en épistémologie, ou en histoire des sciences. L'objectif est double, d'une part caractériser finement les évolutions des communautés scientifiques, d'autre part, à une échelle de temps plus importante, observer les mutations opérées dans la dynamique d'évolution des sciences afin de caractériser des transitions sur le régime de

[1] "a paradigm is what the members of a scientific community share, and, conversely, a scientific community consists of men who share a paradigm"

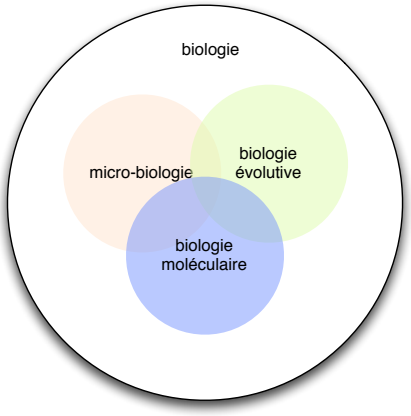


Figure 2: Exemple schématique de l'organisation d'un champ. Les trois sous-champs représentés sont tous ici de la biologie, mais leur intersection est non-nulle

fonctionnement et de régulation même des communautés scientifiques. Est-il possible de repérer une tendance à la balkanisation des sciences associée à l'internationalisation des communautés ? Si Internet n'a, a priori, pas vocation à modifier l'ordre social des communautés scientifiques, il peut affecter le mode de production des savoirs en offrant des moyens d'échanges, de collaboration et de partage des données inédits. Des outils de suivi des dynamiques scientifiques observées in-vivo pourraient nous informer sur ces mutations.

Une autre contrainte de notre travail est de rendre compte du caractère essentiellement multi-échelle de la connaissance scientifique. Par exemple, les universités opèrent classiquement une division des sciences en grands départements qui correspondent à autant de disciplines comme la biologie, l'économie, l'informatique, la physique, etc... Chacune de ces disciplines peut par la suite être morcelée en sous-champs: biologie végétale, animale, moléculaire, évolutive (voir figure 2)... Naturellement les frontières entre champs ne sont pas parfaitement hermétiques et nombre de ces communautés se recouvrent. De plus les séparations entre communautés scientifiques ne suivent pas nécessairement les même lignes de démarcation selon que l'on tâche de différencier l'objet d'étude (biologie animale/végétale) ou le niveau d'approche (micro-biologie/physiologie/écologie) par exemple. Dans la plupart des cas, un sous-champ est spécifique d'un seul champ disciplinaire, mais dans certains cas un sous-champ est précisément défini comme l'intersection de plusieurs champs (la bio-physique par exemple). La structure générale que nous aimerions pouvoir mettre en évidence n'est donc pas celle d'un arbre dont les branches s'affinent en s'allongeant, mais plutÙt celle d'un treillis qui autorise des ramifications de ses branches.

L'objectif que nous nous fixons est de reconstruire la hiérarchie propre à l'organisation des sciences en respectant la complexité des motifs et de leurs articulations à l'aide d'outils d'analyse quantitative et à partir de la simple connaissance de statistiques de base sur les occurrences et cooccurrences d'un ensemble de termes extraits d'un corpus de publications scientifiques.

Dans un premier temps, nous allons introduire une mesure de proximité entre termes qui permet de prendre en compte l'hétérogénéité des fréquences d'utilisation des termes. Nous

exposerons les principales propriétés de cette mesure de proximité et ses principaux avantages par rapport aux mesures employées classiquement en bibliométrie et en scientométrie . Les méthodes de catégorisation permettant de reconstruire la structure multi-échelle et recouvrante des sciences seront également présentées. Puis nous proposerons une méthode de reconstruction des dynamiques de champs.Un cas d'étude sera plus précisément évalué dans une dernière partie ainsi que les perspectives ouvertes par la systématisation de ces méthodes.

# 2 Reconstruction et cartographie

## 2.1 jeux de données

Nos méthodes de reconstruction et de représentation des dynamiques scientifiques seront appliquées à deux jeux de données. Le premier a trait au champ des systèmes complexes, il est constitué d'un corpus de près de $450$ termes, dont on a extrait le nombre de cooccurrences (dans le texte intégral des articles)observées annuellement dans la base de données *scirus* de 1975 à 2005. La base originale est composée de plus de $20.000.000$ publications couvrant un large éventail de contenus scientifiques[2]. La seconde base de données traite de la biologie contemporaine exposée aux évolutions paradigmatiques introduites par l'introduction de la "pensée réseau". Cette dernière sera détaillée plus précisément dans la dernière section.

## 2.2 Mesurer la distance entre termes

La détection de la structure formée par la construction d'un réseau de cooccurrences sur un ensemble de termes est un des principaux objectifs des études de la scientométrie. Doyle a remarqué le premier que la navigation dans les grandes base de données scientifiques était rendue inefficace en raison du manque de pertinence des modes de recherche traditionnelle par mots-clés [7]. L'analyse par "co-présence de termes" ( "co-word analysis") a tâché de répondre à ce constat [3, 4, 16] en introduisant un indice de similarité entre termes qui s'exprime pour deux termes $i$ et $j$ comme le ratio entre le nombre de cooccurrences des termes $i$ et $j$ avec le produit des occurrences de chaque terme.

Cette indice de similarité est symétrique: étant donné un terme $i$ et un autre terme $j$, $i$ sera à la même distance de $j$ que $j$ de $i$. Cette propriétés peut s'avérer problématique lorsqu'on mesure des termes dont les fréquences d'apparition sont très différentes, une mesure symétrique étant aveugle à cette hétérogénéité. Si l'on considère la distance qui sépare un terme $i$ à deux termes $j_1$ et $j_2$, $j_1$ apparaissant aussi fréquemment que $i$ et ayant un recouvrement faible avec $i$ $j_2$ étant beaucoup plus rare mais apparaissant systématiquement avec $i$. Une mesure classique ne permet pas de différencier $j_1$ et $j_2$ du point de vue de $i$. L'utilisation de telles métriques induit un aplanissement des relations entre termes qui se reflète dans les cartes construites.

Afin de rendre compte de l'hétérogénéité de la distribution des fréquences dans un corpus, nous avons proposé[5] une mesure alternative appelée *proximité paradigmatique* et qui

[2]ScienceDirect, Society for Ind. & App. Mathematics, BioMed Central, Crystallography Journals Online, Institute of Physics Publishing, MEDLINE/PubMed, Project Euclid, Scitation and Pubmed Central.

définit la similarité entre les termes $i$ et $j$ comme suit:

$$\mathcal{S}_t^\alpha(i,j) = (n_{ij}^t/n_i^t)^{1/\alpha}(n_{ij}^t/n_j^t)^\alpha$$

o $n_i^t$ et $n_j^t$ désignent le nombre d'occurrences de $i$ et $j$ observée au temps $t$ et $n_{ij}^t$ correspond au nombre de co-occurrences de $i$ et $j$. Le paramètre de focus $\alpha$ est un paramètre de la mesure réel et positif . Cette proximité paradigmatique a les propriétés suivantes (dans un souci de clarté, les paramètres $\alpha$ et $t$ ont été omis ):

1. $\mathcal{S}(i,j) = 0$ si $n_{ij} = 0$
2. $\lim_{\frac{n_{ij}}{n_i} \to 0}(\mathcal{S}(i,j)) = 0$
3. $\mathcal{S}(i,i) = 1$
4. $\mathcal{S}(i,j)$ est croissant lorsque $n_{ij}$ croît, toute choses étant constantes par ailleurs. A contrario, une augmentation de $n_i$ ou $n_j$, $n_{ij}$ restant constant, entraîne une diminution de la valeur de $\mathcal{S}(i,j)$
5. Dans l'hypothèse d'un échantillon représentatif, $\mathcal{S}(i,j)$ est indépendant du nombre total d'articles dans la base de données.

Cette mesure vérifie également la propriété suivante: $\mathcal{S}^\alpha(i,j) = \mathcal{S}^{1/\alpha}(j,i)$. Elle peut s'interpréter géométriquement comme une mesure de quasi-inclusion. $\alpha \to 0$ alors $\mathcal{S}(i,j) > 0 \iff n_{ij}^t = n_j^t$ (dans une vision ensembliste $I \in J$ si on désigne par $I$ (respectivement $J$) l'ensemble des articles mentionnant $i$ (resp. $j$)) et symétriquement: $\alpha \to \infty$, $\mathcal{S}(i,j) > 0 \iff n_{ij}^t = n_i^t$ ($J \in I$).

Notre proximité paradigmatique permet de définir le voisinage d'un terme cible $i$ étant donné un seuil $s$, un focus $\alpha$ et un temps $t$ comme:

$$V_{s,\alpha}^t(i) = \{j | \mathcal{S}_t^\alpha(i,j) > s\}$$

Comme nous le préciserons dans la section suivante, le paramètre de focus permet "d'orienter" la recherche de termes voisins soit vers des termes de même importance ($\alpha = 1$) soit des termes plus spécifiques que le terme cible original ($\alpha > 1$), soit des termes plus généraux ($\alpha < 1$).

### 2.3 vers une cartographie multi-niveau des sciences

Un objectif classique en bibliométrie est de produire des cartes des paysages conceptuels observés à travers des bases d'articles scientifiques[2, 14]. De nombreuses méthodes de clustering comme les cartes de Kohonen ont été utilisées pour faciliter la visualisation de très grandes bases de données en tâchant d'en extraire les principaux champs de recherche [13, 20].

Nous allons exposer notre propre méthode de navigation et de clusterisation qui exploite pleinement les propriétés de notre mesure entre termes. Nous proposons également

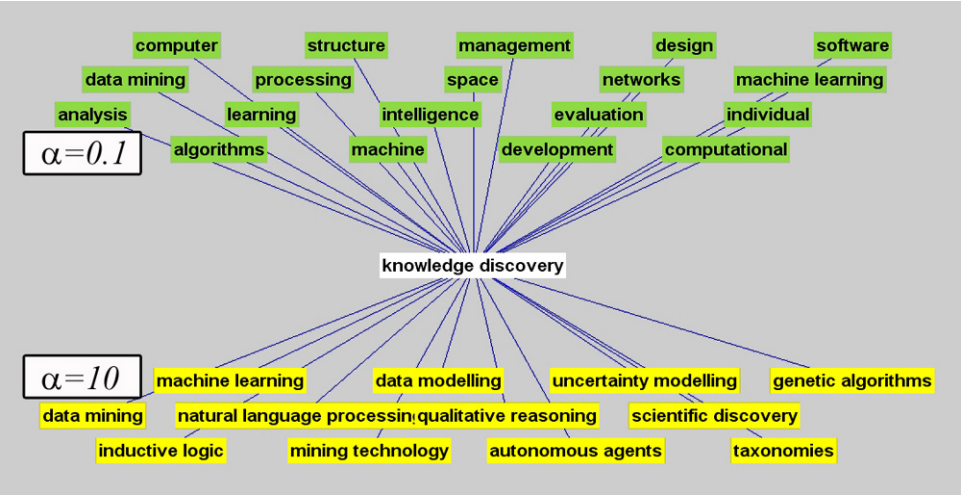


Figure 3: Voisinages en spécificité et en généricité du terme *knowledge discovery*

de fournir une représentation d'un corpus de termes à trois niveaux différents. Dans un premier temps, nous définissons un niveau microscopique donnant accès aux voisinages locaux des termes. Puis, le niveau mésoscopique est construit à partir de la matrice de distances inter-termes, ce niveau permet de retrouver les domaines scientifiques pertinents constitués de sous-ensembles de concepts. Enfin le niveau macroscopique est construit à partir du niveau mésoscopique en employant la même méthode de détection d'ensembles cohérents que pour le passage micro-macro.

**échelle microscopique: voisinages paradigmatiques** Cette première approche est purement locale. Etant donné un terme cible, nous cherchons à identifier ses voisins les plus proches. Le paramètre $\alpha$ permet d'accéder à deux types de voisinage. Pour de faibles valeurs de $\alpha$ ($\alpha < 1$), les plus proches voisins ont tendance à avoir un caractère plus génériques que le terme cible (les ensembles $J$ qui englobent $I$ sont avantagés par la mesure). Si le paramètre $\alpha$ est plus important ($\alpha > 1$), on retrouvera préférentiellement des termes plus spécifiques que $i$ ($J$ inclus dans $I$). La figure 3 extraite d'un cas d'étude sur un corpus de termes portant sur les sciences cognitives illustre cette propriété. Nous avons tracé le voisinage $V_{s,\alpha}$ du terme *knowledge discovery* pour $\alpha = 0,1$ et $\alpha = 10$ et une valeur de seuil $s$ fixée. Pour $\alpha = 10$ les termes les plus proches de *knowledge discovery* le spécifient en fournissant les termes utilisés dans les sous-spécialités du domaine (dans l'exemple figure 3 "natural language processing", "scientific discovery, etc..."). A contrario, un paramètre de focalisation plus petit $\alpha = 1/10$ a tendance a entourer le terme cible de l'ensemble de ses contextes (dans notre exemple: "algorithm", "processing", "learning", etc...)

Dans le cas particulier o $\alpha = 1$ on exclut les termes trop génériques et trop spécifiques de notre voisinage pour sélectionner préférentiellement des termes de même fréquence dans le corpus. La proximité paradigmatique est alors égale à l'indice d'équivalence (e-coefficient) introduit par Callon [3].

**méso-échelle: identification des domaines paradigmatique** Si l'on examine la partie inférieure de la figure 3, on observe que plusieurs domaines partageant la proximité avec le terme *knowledge discovery* semblent co-exister . Une partie de ces termes est orientée vers les outils d'apprentissage ("machine learning") tandis qu'une autre est centrée sur la fouille de données ("data mining"). Pour détecter de faÁon automatique ces nuances, il faut faire appel à des contextes plus larges que les simples informations locales de proximité deux à deux. L'objectif est d'exploiter les informations sur l'ensemble des relations entre termes pour extraire de faÁon bottom-up les différentes pratiques desquelles peuvent relever un terme donné. On cherche donc à classer automatiquement les données en fonction des valeurs de la proximité paradigmatique $\mathcal{S}^{\alpha}$ calculées entre chaque paire de termes.

La littérature sur les algorithmes de détection de communautés dans les réseaux est pléthorique (pour un examen de quelques méthodes récentes et leur évaluation voir [6]), les plus récents d'entre eux visent à effectuer la meilleure partition possible des noeuds d'un réseau en tâchant d'optimiser un facteur de qualité appelé la modularité [15]. Une simple partition de l'ensemble des termes ne permet de dégager que des structures de type purement binaire (un terme est classée de faÁon non ambiguÎ dans une unique catégorie) sous la forme d'arbres. Aussi, une telle méthode ne permet pas de rendre compte de la multiplicité des usages que peut prendre un terme. C'est pourquoi nous préférerons employer un algorithme de "détection de communautés" qui autorise un certain taux de recouvrement entre champs. Ainsi un terme polysémique devrait pouvoir être distribué sur l'ensemble de ses usages possibles en fonction des contextes privilégiés auxquels il est fréquemment associé. Nous souhaitons donc que notre algorithme de détection de champs soit en mesure de classer un terme dans plusieurs groupes différents s'il est susceptible de prendre des sens variés ou simplement d'être employé par des communautés différentes. Plusieurs algorithmes ont été récemment proposés à cette fin, une méthode déterministe remplissant l'ensemble des critères recherchés est la méthode de détection de communautés par percolation de cliques récemment introduite par Palla et al. [18].

A partir de notre mesure de proximité paradigmatique entre termes, nous construisons un réseau lexical qui relie le terme $i$ au terme $j$ si et seulement si $j \in V_{s,\alpha}(i)$ étant donné un seuil $s$ fixé. Muni de ce réseau lexical, on appliqué l'algorithme de percolation de k-clique (k désignant la taille de la clique) dans sa version orientée [19] qui nous permet de décrire les ensembles de termes les plus cohérents.

Une fois les champs extraits, nous proposons de les plonger dans un espace bidimensionnel. Etant donné un champ $C$ et un terme $w$: on définit l'indice de généricité $i_g$ et l'indice de spécificité $i_s$ comme suit:

**indice de spécificité** Il fournit une mesure du positionnement du terme $w$ vis-à vis de l'ensemble de ses voisins dans le champ. C'est la somme des distances entrantes des termes de $C$ vers $w$

$$i_s(w) = \frac{1}{card(C)} \sum_{w' \in C} \mathcal{S}^{\alpha}(w', w)$$

**indice de généricité** Il définit dans quelle mesure le groupe $C$ est un bon voisinage pour le

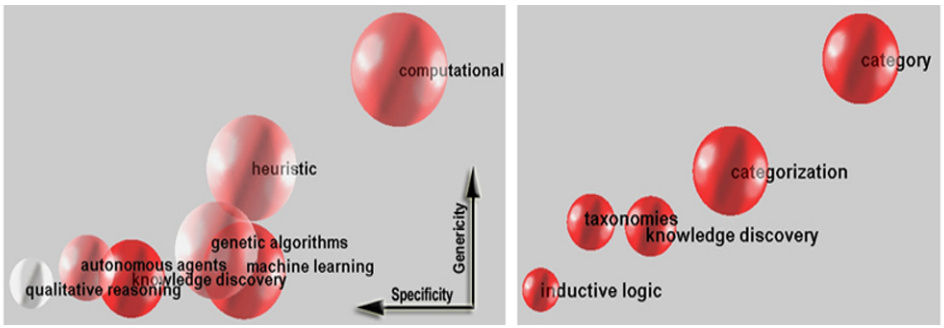


Figure 4: **Two paradigmatic fields mentioning the term *Knowledge Discovery* during the period 2002-2005**. *Knowledge Discovery* appartient à deux sphères de production de connaissance distinctes. A gauche une orientation *machine learning*, à droite l'accent est mis sur les questions de *categorization*. Au sein d'un champ, $i_s$ décroît de gauche à droite,et $i_g$ décroît de haut en bas.

terme $w$. On le définit comme la moyenne des distances sortantes:

$$i_g(w) = \frac{1}{card(C)} \sum_{w' \in C} \mathcal{S}^{\alpha}(w, w')$$

Ces deux indices permettent de représenter de faÁon intuitive les champs dans un espace à deux dimensions. A chaque terme, on attribue une coordonnée $(i_s, i_g)$ et une taille proportionnelle à son importance dans le champ ( et plus précisément à la somme du nombre de ses co-occurrences avec tous les autres termes du champ). La couleur de chaque terme traduit le taux de croissance de son importance dans le champ entre deux périodes consécutives (du blanc: croissance nulle au rouge foncé, croissance supérieure à $50\%$).

Pour l'illustrer, nous présentons figure 4 deux domaines qui partagent les termes "knowledge discovery" au cours de la période 2003-2005. Comme mentionné plus haut, ce terme peut relever de plusieurs domaines distincts, un premier orienté *apprentissage automatique* le second plus focalisé sur les enjeux propres à la *catégorisation* (cf figure 4).

Il convient de souligner ici que cette méso-échelle de visualisation est complémentaire mais distincte du niveau microscopique de visualisation. Les champs détectés regroupent des termes qui vérifient des critères de relation globaux sur l'ensemble de leurs éléments. D'autres exemples de reconstruction automatique de champs sont consultables à l'adresse suivante: http://cssociety.org/CSM.

Compte tenu de la définition de l'ensemble des champs il est désormais possible d'établir une carte dont l'unité de base soit le champ et qui condense l'information à un niveau macroscopique.

**échelle macroscopique**

La prochaine étape consiste désormais à donner un aperÁu de l'articulation des différents champs paradigmatiques que nous avons identifiés à l'échelle méso afin de fournir une vision globale et structurée du paysage scientifique défini par notre ensemble de termes. Etant

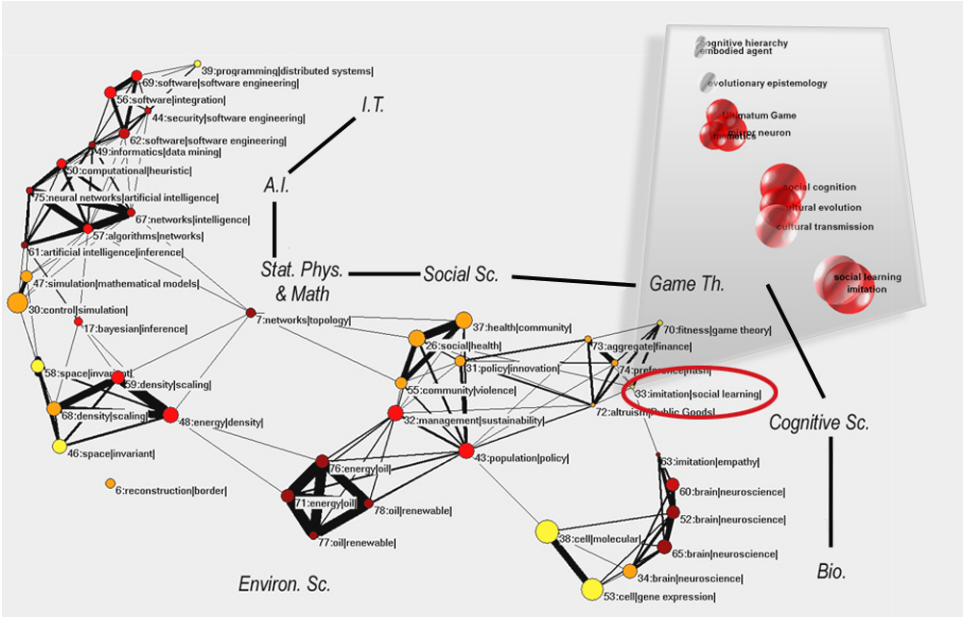


Figure 5: Carte macroscopique du champ des "Systèmes complexes" Les couleurs sombres correspondent aux champs dotés des taux de croissance les plus importants. Chaque champ peut être "déplié" dans son référentiel bidimensionnel comme l'illustre l'insert sur le champ "imitation & social leaning".

donnée une période temporelle, nous avons défini les domaines paradigmatiques comme des ensembles de termes, ces termes pouvant appartenir à plusieurs domaines paradigmatique différents. Une procédure possible pour représenter la faÁon dont ces champs se structurent est de définir un réseau dont les noeuds correspondent aux champs et dont les liens sont définis conformément au recouvrement entre champs. D'autres moyens de définir un lien entre deux champs sont envisageables, notamment en calculant la moyenne des distances deux à deux entre l'ensemble des termes constituant deux champs donnés. Cependant nous nous contenterons dans une première approche de définir le poids d'un lien entre deux champs comme le nombre de termes partagés par ces deux domaines. Les indices de spécificité et de généricité: $i_s$ et$i_g$ sont néanmoins utiles pour définir une fonction de labellisation efficace et informative des champs (dans l'exemple figure 5 nous avons choisi d'étiqueter les champs par leur deux termes les plus génériques).

A titre d'exemple, la figure 5 est une représentation macroscopique d'un corpus de termes associées au domaine des "systèmes complexes" pour la période 2002-2005. La taille d'un nœud est proportionnelle à l'importance du champ $p_i$ (échelle logarithmique) Nous pouvons également visualiser sur cette carte la croissance de l'activité de chaque champ paradigmatique. Pour ce faire, nous calculons pour chaque champ la croissance des cooccurrences de termes au sein du champ $C$ entre une période $T$ et la période précédente $T_{-}$:

$$A_C = \frac{1}{card(C)} \sum_{i \in C} \frac{p_i^T}{p_i^{T_-}}$$

o $p_i^T$ est défini comme $p_i^T = \frac{n_i^T}{\sum_j n_j^T}$ Les champs de couleur bleue ont un taux de croissance négatif, jaune, rouge, et brun un taux de croissance positif et d'autant plus fort que la couleur s'assombrit. Une carte interactive qui permet de zoomer dans les endroits d'intérêt et de naviguer à travers les champs paradigmatique peut être consultée en ligne à l'adresse suivante: http://cssociety.org/CSM.

# 3 Méthode de reconstruction dynamique

L'analyse statique de la structure des champs paradigmatiques est une première étape vers la caractérisation et la représentation des dynamiques propres à l'évolution des sciences. L'ensemble des méthodes décrites dans la section précédente repose sur un corpus daté, et permettent donc de reconstruire la structure d'un domaine scientifique à n'importe quelle période. On peut ainsi aisément imaginer rajouter une dimension temporelle à notre analyse. Ce second volet ouvre la voie à de nombreuses questions et applications [1, 8]. Est-il possible de reconstruire l'évolution des changements paradigmatiques majeurs, peut-on identifier de faÁon automatique les approches émergentes ? Peut-on également retracer les grandes mutations dans les régimes de régulation et d'organisation des communautés scientifiques...? Plusieurs approches sont envisageables, d'abord au niveau microscopique puis au niveau mésoscopique.

## 3.1 dynamiques de voisinage

Au niveau local, on peut s'interroger sur l'évolution des voisinages associés à un terme. Etant donné un seuil $s$ fixé et un terme $i$ on peut représenter l'ensemble des termes qui appartiennent au voisinage de $i$ à différentes périodes. Cette représentation offre une première faÁon d'observer le glissement de sens d'un terme au cours du temps comme illustré figure 6.

## 3.2 dynamiques paradigmatiques

Une première approche naÔve des dynamiques scientifiques consisterait à mettre bout à bout l'ensemble des cartes macroscopiques obtenues à des périodes successives. Une telle frise temporelle (cf. figure 7) n'est informative d'un point de vue dynamique que sur la structuration globale des champs. Observe-t-on une augmentation ou une diminution du nombre total de champ, est-ce que la cohésion de l'ensemble de ces champs a tendance à augmenter ou à diminuer, etc...? Malheureusement rien ne permet a priori de décrire une dynamique fine sur les champs paradigmatiques eux-mêmes. Pour ce faire il faudrait être capable, à la manière de notre œil qui identifie des structures remarquables et est capable d'en prédire les déformation et les mouvements, d'identifier de manière non ambiguÎ la faÁon dont un champ paradigmatique à un moment donné se transforme en un autre champ à un moment ultérieur [21, 17]. Cette problématique revient à postuler une forme de stabilité entre deux périodes de temps, stabilité qu'il faut alors définir et quantifier.

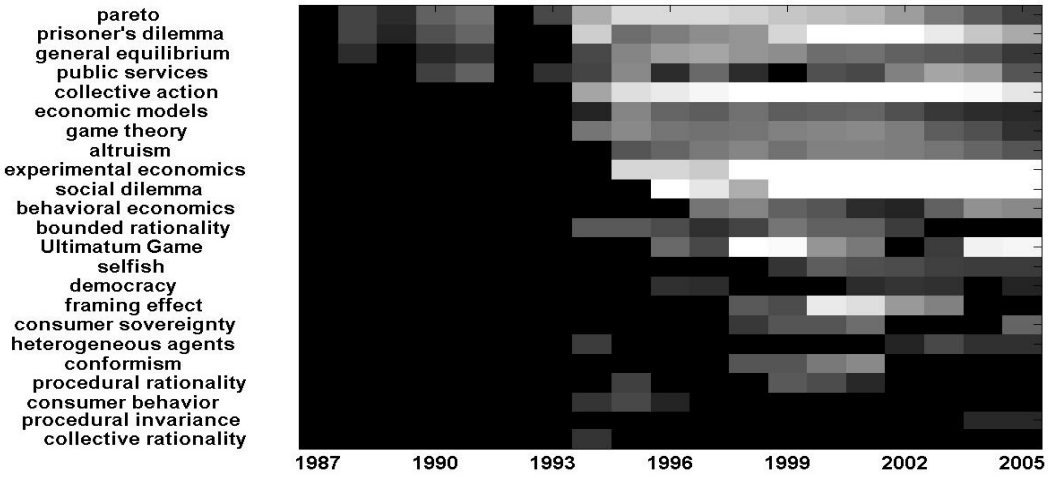


Figure 6: Représentation dynamique de l'évolution du voisinage du terme "Public Good" de 1987 à 2005 pour $\alpha = 1$. Une zone noire signifie que le terme associé n'est pas dans le voisinage de "Public Good" une année donnée. Les cases les plus claires correspondent par contre aux voisins les plus proches. On observe sur cet exemple que les études sur les biens publics ont été appréhendées récemment par une approche de théorie des jeux. Parmi les termes émergeant dans le voisinage de "Public Good" on trouve notamment "heterogeneous agents" ou "procedural rationality". Cette dynamique correspond bien aux transformations actuelles subies dans ce domaine.

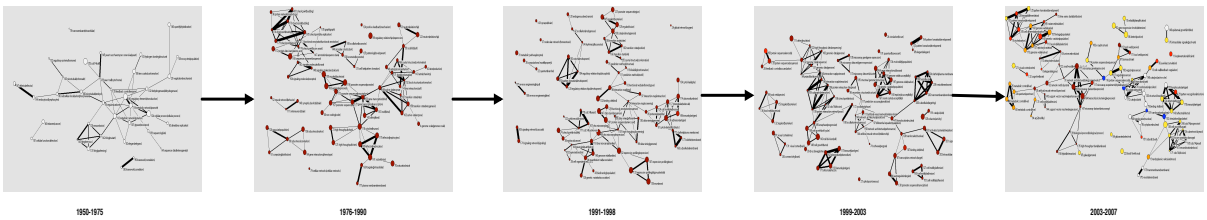

Figure 7: Evolution de la structuration des champs de la biologie & réseaux sur cinq périodes: 1950-1975 ; 1976-1990 ; 1991-1998 ; 1999-2002 ; 2003-2007 .

Nous proposons une procédure simple pour réaliser cet appariement inter-temporel entre communautés. L'argument de départ est simple. Etant donné un champ, on cherche à détecter le champ ou les combinaisons de champs qui sont les plus ressemblants et donc les plus probablement appariés. Nous définissons donc un critère simple sur l'ensemble des pères possibles d'une communauté $C_i^{t+1}$ au temps $t$. Les champs "pères" de $C_i^{t+1}$ sont définis comme l'ensemble $\mathcal{F}_i^{t+1}$ dans l'ensemble des parties des champs paradigmatiques à la période $t$ qui sont les plus semblables au champs fils. Par mesure de simplicité on choisit une distance de Jaccard classique pour mesurer cette similarité. Ainsi on définit l'ensemble

des fils du champ $i$ au temps $t+1$ comme:

$$\mathcal{F}_i^{t+1} = \underset{K \in \mathcal{P}(C^t)}{\operatorname{argmax}} \frac{|C_i^{t+1} \bigcap (\cup_{j \in K} C_j^t)|}{|C_i^{t+1} \bigcup (\cup_{j \in K} C_j^t)|}$$

La procédure d'identification est illustrée figure 8. Etant donnés deux champs $A$ et $B$ à une période $t1$, on détecte à un moment ultérieur $t2$, deux autres champs $C$ et $D$ (qui se recouvrent en un noeud). La question est donc de savoir quel est l'ascendance la plus crédible des champs $C$ et $D$. Compte tenu de notre définition, on voit immédiatement que la communauté $C$ descend directement de la communauté $A$, même si deux noeuds ont disparu, tandis qu'un autre a été ajouté, la distance entre les champs $A$ et $C$ vaut: $d(A, C) = \frac{2}{5}$ et constitue la meilleure correspondance possible. En ce qui concerne $D$, les choses sont un peu moins simples, mais il suffit alors de lister les différents cas possibles: selon que le champ $D$ descende de $A$, $B$ ou $A \cup B$, on calcule les ratios suivants: $d(A, D) = \frac{2}{8}$, $d(B, D) = \frac{3}{6}$ et enfin $d(A \cup B, D) = \frac{5}{7}$. La meilleure correspondance possible pour le champ $D$ est donc offerte par $A \cup B$.

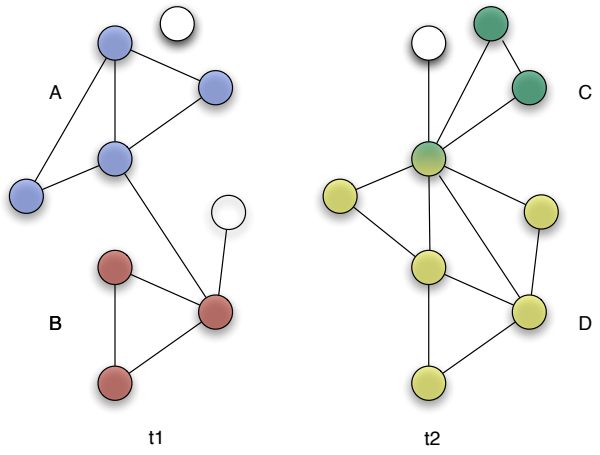


Figure 8: Comparaison inter-temporelle de motifs.

## 4 Etude de cas: la biologie contemporaine à l'épreuve des réseaux

Cette méthode de reconstruction des dynamiques scientifiques a été appliquée à un corpus de termes et de publications en biologie afin de tenter d'éclairer l'histoire récente des renouvellements paradigmatiques ouverts 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. Ce travail est réalisé en collaboration avec des historiens des sciences (Christophe Bonneuil et Jean-Paul Gaudillère).

### 4.1 délimitation du corpus

La sélection du corpus de termes a été réalisée en deux étapes. Dans un premier temps, un ensemble de publications a été sélectionné à partir d'une requête sur l'*ISI Web of Knowledge* portant sur un ensemble de journaux de premier rang dont le sujet devait contenir le terme "network"[3]. Cette première extraction a permis de construire une liste de termes caractéristiques extraits des abstracts de cette collection d'articles. Un ensemble de plus de 800 termes a ainsi été sélectionné. Une fois le corpus de termes défini, une matrice de cooccurrences des termes a été construite à partir de la base de données *pubmed* sur les cinquante dernières années afin d'illustrer les mutations profondes subies par la biologie sur cette période. La figure 9 fournit une représentation macro de ce corpus sur la période 2004-2007.

### 4.2 reconstruction statique

Les cartes produites par les méthodes de reconstruction des dynamiques scientifiques peuvent ainsi 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 7). Les cartes ainsi produites peuvent servir de support à l'exploration d'hypothèses expliquant les mutations épistémologiques contemporaines.

La carte figure 9 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 rÙle prépondérant des outils et instruments 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 faÁon superficielle au coeur instrumental dur, même si on peu 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, algorithme génétique...). Les approches écosystémiques sont également très actives, et voient leur composition évoluer (champs 77, 112: présence de réseaux trophiques).

[3]la requête précise est la suivante: (TS=Network*)AND (SO=("Science" OR "Nature" OR "Proceeding of the National Academy of Science" OR "Nature Genetics" OR "Annual Review of Genetics" OR "Annual Review of Biochemistry" OR "Annual Review of Cell and Developmental Biology" OR "Annual Review of Genomics and Human Genetics" OR "Journal of Theoretical Biology" OR "Biochimica et Biophysica Acta" OR "Nucleic Acids Research" OR "Journal of Molecular Biology" OR "Genetics" OR "Current Biology" OR "Genome Research" OR "Genome Biology" OR "Bioinformatics" OR "Biosystems" OR "BMC Systems Biology"))

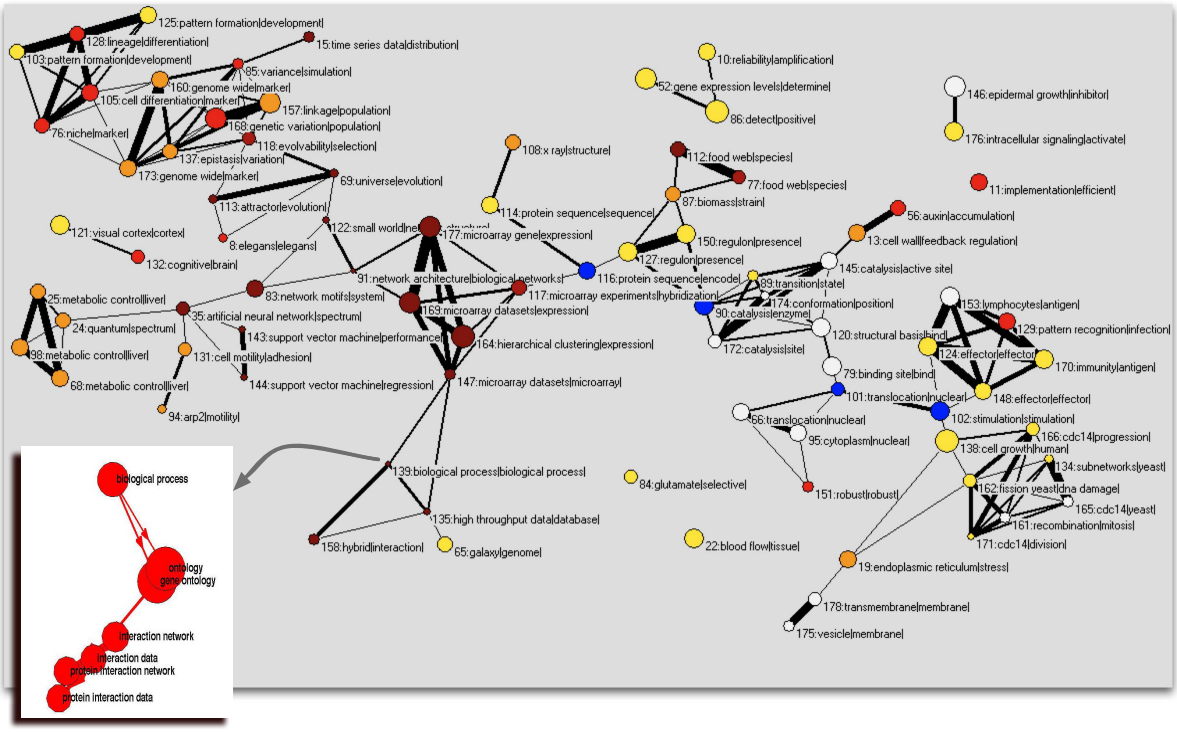


Figure 9: paysage conceptuel de la période 2004-2007 autour du thème biologie et réseau. Chaque noeud du réseau représente un champ paradigmatique illustré par un exemple dans l'encart en bas à gauche. La taille d'un champ est proportionnelle à son activité tandis que sa couleur reflète la croissance de son taux d'activité (bleu: décroissance de plus de $50\%$, blanc, champ stable, jaune, croissance de $50\%$, rouge, croissance supérieure à $100\%$.

### 4.3 analyse des dynamique

Dans l'exemple illustré figure 10, on a tracé la phylogénie des champs depuis 1963 en surlignant 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, et les périodes charnières les plus riches en termes de nouveaux champs par exemple. Ce type de représentation doit être interprété avec prudence. En effet, l'absence de descendance d'un champ à une période donnée ne signifie pas que le champ ait complètement disparu mais qu'il est moins actif que précédemment. En effet, l'ensemble des champs détectés ont été calculés avec l'algorithme de percolation de cliques qui dépend directement du réseau lexical construit à partir des statistiques de cooccurrences. Or ce graphe lexical peut être plus ou moins dense selon la valeur de seuil $s$ choisie pour calculer les voisinages. Ainsi, en abaissant ce seuil de nouveaux champs pourraient apparaître et certaines transitions entre périodes seraient susceptibles d'être modifiées. La représentation actuelle ne fait figurer que les champs dont l'activité est suffisamment importante pour être détectés comme des structures pertinentes. Une carte plus exhaustive, plutÙt que de faire figurer les seuls champs dont l'activité est supérieure à

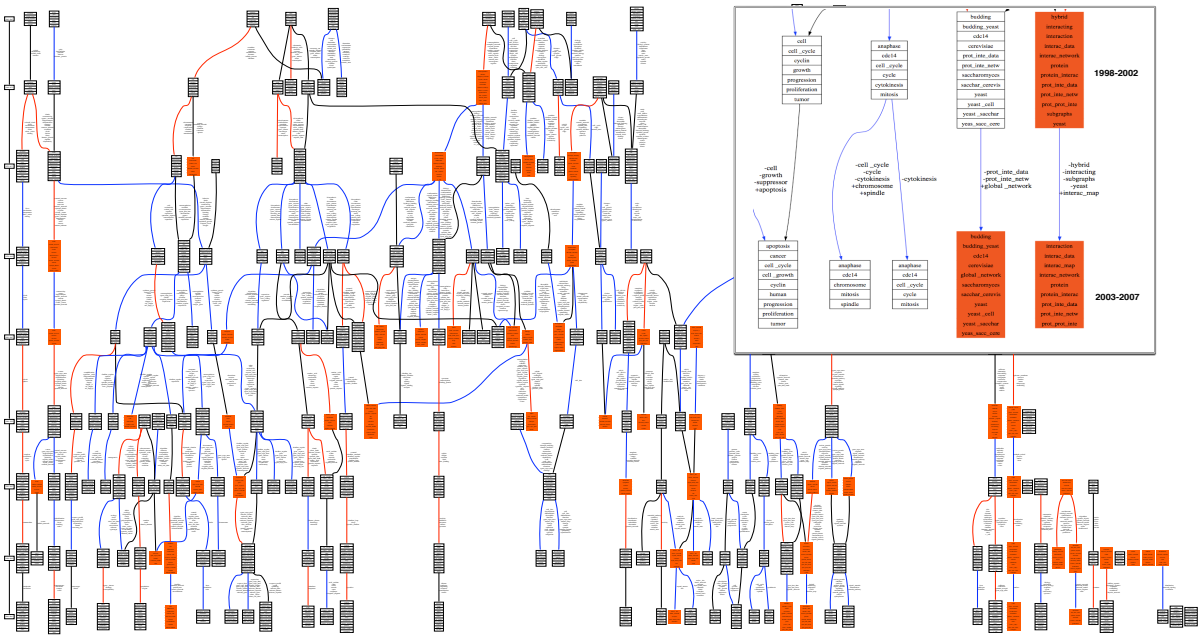


Figure 10: Phylogénie des champs paradigmatiques de 1963 à 2007. L'ensemble des champs contenant le terme "réseau" sont en rouge. Sur une ligne on retrouve l'ensemble des champs et leur composition sur une période de quatre ans. Un lien bleu signifie qu'il y a eu globalement décroissance de la communauté, un lien rouge, croissance (enrichissement du nombre de termes). Les légendes le long des liens signalent les termes acquis et perdus d'une période à la suivante. Les événements possibles sont: naissance (absence de père), mort (absence de fils), croissance (un père, lien rouge), décroissance (un père, lien bleu), scission (une père plusieurs fils ), fusion (un fils, deux pères, liens rouges & noirs ou bleus & noirs). En haut à droite agrandissement sur les deux dernières périodes.

un seuil fixé, pourrait faire figurer l'ensemble des champs associés à un indice d'activité qui informe sur la force du champ en question à une période donnée. C'est pour cette raison que ce type de reconstruction n'a pas encore donné lieu à une interprétation de la part des experts.

## Perspective

Les outils développés ont 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 de connaissance multi-échelle sont ici cruciales autant à cause de la complexité des données à mettre en forme que par la nécessité d'intégrer des experts du champ dans la boucle de modélisation. L'aspect mutli-niveau des cartes a été pris en compte en développant un site web, permettant de naviguer à travers les différentes périodes et à travers les différents niveaux selon le degré de résolution souhaitée. La dimension dynamique est un challenge supplémentaire qui requiert d'inventer des solutions de

visualisation inédites, le treillis des champs et de leurs transitions n'étant manifestement pas une solution satisfaisante pour le moment.

## Conclusion

Nous avons essayé de montrer la manière dont la cartographie des science pourrait bénéficier d'une mesure asymétrique de proximité entre termes. Une méthode de catégorisation avec recouvrement nous a permis de reconstruire une structure hiérarchisée des sciences qui soit robuste à la polysémie des termes et aux enchâssements complexes des communautés scientifiques. Ces méthodes de reconstruction ouvrent la voie à de nouveaux modes de navigation dans les bases de données qui pourraient s'avérer utiles pour les chercheurs, les historiens des sciences, mais aussi les responsables de politiques scientifiques. De plus ces méthodes ne se restreignent pas nécessairement au monde scientifique. Toute autre base de textes est susceptible d'être traitée selon la même procédure (brevets, presse, contenus en ligne, etc...).

## Remerciements

Ce papier a été développé dans le cadre du projet COBINA soutenu par le programme OGM de l'ANR. Il a bénéficié de l'expertise précieuse de Christophe Bonneuil et Jean-Paul Gaudillère, des conseils avisés de Pierre-Benoît Joly ainsi que du concours de David Chavalarias.

## References

[1] Robert R Braam, Henk F Moed, and Anthony F J van Raan. Mapping of science by combined cocitation and word analysis. ii. dynamical aspects. *Journal American Society Information Science*, 42(4):252–266, 1991.

[2] R K Buter and E C M Noyons. Using bibliometric maps to visualise term distribution in scientific papers. pages 697–702, 2002.

[3] M Callon, J P Courtial, and F Laville. Co-word analysis as a tool for describing the network of interaction between basic and technological research: The case of polymer chemistry. *Scientometric*, 22(1):155–205, 1991.

[4] Michel Callon, J Law, and A Rip. *Mapping the dynamics of science and technology*. 1986.

[5] David Chavalarias and Jean-Philippe Cointet. Bottom-up scientific field detection for dynamical and hierarchical science mapping - methodology and case study. *Scientometric*, 75(1), 2008.

[6] L Danon, A Diaz-Guilera, J Duch, and A Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, Jan 2005.

[7] Lauren B Doyle. Semantic road maps for literature searchers. *J. ACM*, 8(4):553–578, 1961.

[8] Eugene Garfield. Historiographic mapping of knowledge domains literature. *Journal of Information Science*, 30(2):119–145, 2004.

[9] Francis HEYLIGHEN, Margeret HEATH, and Frank VAN OVERWALLE. The emergence of distributed cognition: a conceptual framework. *Proceedings of Collective Intentionality IV, Siena (Italy)*, Oct 2004.

[10] D Hull. *Science as a process: an evolutionary account of the social and conceptual development of science*. 1988.

[11] Thomas S Kuhn. *The Structure of Scientific Revolutions, Postscript*. 1969.

[12] Bruno Latour. *Science in Action: How to Follow Scientists and Engineers Through Society*. Oct 1988.

[13] X Lin and D Soergel. A self organizing semantic map for information retrieval. *Proc. 14th International SIGIR Conference*, pages 262–269, 1991.

[14] Irina Marshakova-Shaikevich. Bibliometric maps of field of science. *Infometrics*, 41(6):1534–1547, 2005.

[15] M Newman and E Leicht. Mixture models and exploratory data analysis in networks. *Arxiv preprint physics*, Jan 2006.

[16] ECM Noyons and AFJ van Raan. *Dealing with the data flood. Mining data, text and multimedia.*, pages 64–72. 2002.

[17] Gergely Palla, Albert-Laszlo Barabási, and T Viscek. Quantifying social group evolution. *Nature*, Jan 2007.

[18] Illés J Farkas Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–8, Jun 2005.

[19] Illés J Farkas Palla, I Farkas, P Pollner, I Derenyi, and T Vicsek. Directed network modules. *New Journal of Physics*, Jan 2007.

[20] Yao Sun. Methods for automated concept mapping between medical databases. *J. of Biomedical Informatics*, 37(3):162–178, 2004.

[21] Yi Wang. Community evolution of social network: Feature, algorithm and model. *arxiv*, physics.soc-ph, Apr 2008.