# *DECOUVERTE DE CONNAISSANCES DANS LES BASES DE DONNEES BIBLIOGRAPHIQUES*

## *LE TRAVAIL DE DON SWANSON : DE L'IDEE AU MODELE*

---

**Jean-Dominique Pierret,**
Galderma R&D
jeandominique.pierret@galderma.com + 33 4 93 95 70 48

**Eric Boutin**
Laboratoire LePont
boutin@univ-tln.fr + 33 4 94 14 23 56

**Adresse professionnelle**
Galderma R&D △ 635, route des Lucioles △ BP 87 △ F-06902 Sophia-Antipolis Cedex
Laboratoire LePont △ Université de Toulon-Var △ BP 132 △ F-83957 La Garde Cedex

**Résumé** : On considère que l'information disponible dans les bases de données bibliographiques est une information datée, validée par un processus long qui la rend peu innovante. De plus, dans leur mode d'exploitation, les bases de données bibliographiques sont classiquement interrogées de manière booléenne : le résultat d'une requête est un ensemble d'informations connues qui n'apporte en lui-même aucune nouveauté. Les travaux de Don Swanson montrent le potentiel insoupçonné des bases bibliographiques dans la révélation et la découverte de connaissances. Cet intérêt ne tient pas tant à la nature de l'information disponible qu'à la méthodologie utilisée pour révéler ces nouvelles connaissances. Cette méthodologie générale s'applique de façon privilégiée dans un environnement d'information validée et structurée ce qui est le cas de l'information bibliographique.
L'expression Knowledge Discovery in Databases (KDD) désigne une méthodologie de création de nouveaux savoirs à partir de bases de données bibliographiques.
Dans cet article, nous aborderons successivement le principe général du KDD à partir des travaux de Don Swanson, puis la méthode employée pour découvrir de la connaissance dans les bases de données bibliographiques biomédicales.

**Summary**: It is now considered that the information available in bibliographical databases is dated, validated through a long process which does not make it very innovative. Furthermore, database processing is normally performed using boolean operators : the results obtained from a query provides a sum of expected information which, in itself, does not deliver any novelty. Don Swanson's work demonstrates the unsuspected potential of bibliographical databases in revealing and discovering knowledge.The interest of his approach lies less on the available information itself than on the methodology used to disclose new knowledge. This general

methodology fits perfectly well within an environment of validated and structured information, as is the case for bibliographical data.
The expression Knowledge Discovery in Databases (KDD) indicates a methodology which creates new knowledge based upon bibliographical data.
In this article, we will cover the principals of KDD based on Don Swanson's work as well as the method used to disclose knowledge within biomedical bibliographical databases.

**Mots clés** : découverte de connaissances, bases de données bibliographiques.

**Key words:** knowledge discovery, bibliographic databases.

# Découverte de connaissances dans les bases de données bibliographiques
# Le travail de Don Swanson : de l'idée au modèle

Ce titre comporte un paradoxe. Traditionnellement, en Sciences de l'Information, on considère que l'information disponible dans les bases de données bibliographiques est une information datée, validée par un processus long qui la rend peu innovante. De plus, dans leur mode d'exploitation, les bases de données sont classiquement interrogées de manière booléenne : le résultat d'une requête est un ensemble d'informations connues qui n'apporte en lui-même aucune nouveauté. Les travaux de Don Swanson montrent le potentiel insoupçonné des bases bibliographiques dans la révélation et la découverte de connaissances. Cet intérêt ne tient pas tant à la nature de l'information disponible qu'à la méthodologie utilisée pour révéler ces nouvelles connaissances. Cette méthodologie générale s'applique de façon privilégiée dans un environnement d'information validée et structurée ce qui est le cas de l'information bibliographique.
L'expression « découverte de connaissance dans les bases de données » est la traduction de *Knowledge Discovery in Databases* (KDD, que nous emploierons par la suite) ou *Text-based Knowledge Discovery*. Cette expression peut recouvrir beaucoup de techniques, de manière assez générique, comme le *text mining* ou la classification. Dans cet article, le KDD est pris au sens littéral, c'est-à-dire comme une méthodologie de création de nouveaux savoirs à partir de bases de données bibliographiques.

Don R. Swanson est mathématicien de formation et a manifesté un grand intérêt pour l'information biomédicale. Professeur émérite de l'Université de Chicago, il a reçu la plus haute distinction de l'ASIST[1] en 2000 (ASIST *Award of Merit*) pour l'ensemble de ses travaux sur le KDD.
Dans cet article, nous aborderons successivement le principe général du KDD puis la méthode employée pour découvrir de la connaissance dans les bases de données bibliographiques biomédicales.

## 1 – LA PREMIERE DECOUVERTE : L'HUILE DE POISSON ET LA MALADIE DE RAYNAUD

Au début des années 80, Don Swanson est sollicité par une revue de vulgarisation américaine pour écrire un article sur l'alimentation des esquimaux. La consommation de poissons et de mammifères marins, riches en acides gras poly-insaturés longs, diminue le facteur de risque de maladies cardiovasculaires, d'où leur moindre incidence chez les esquimaux [Dewailly, 2001]. Swanson fait des recherches bibliographiques dans ce sens et il trouve que :

- l'huile de poisson, composée en grande partie de tels acides gras, était connue pour diminuer la viscosité du sang et l'agrégation des plaquettes (favorise la prévention des thromboses et de l'athérosclérose) et pour agir sur la réactivité vasculaire, d'une part,
- et d'autre part, dans la maladie de Raynaud[2] la viscosité du sang et l'agrégation plaquettaire augmentent et se produit une vasoconstriction exagérée.

Le lien est évident et Swanson est le premier à formuler l'hypothèse selon laquelle l'huile de poisson est un traitement potentiel de la maladie de Raynaud. En effet, avant 1986, aucun document ne lie l'huile de poisson et la maladie de Raynaud. Une publication détaille son hypothèse d'un point de vue physiologique [Swanson, 1986] et une autre expose brièvement la méthode employée [Swanson 1987]. En 1989, une équipe de cliniciens (Albany Medical College, New York) montre que même si l'huile de poisson ne permet pas de guérir de la maladie de Raynaud, elle contribue à améliorer l'état des malades [DiGiacomo, 1989].

En 2000, Swanson résume ainsi son processus de découverte : « *In 1985, I was struck by lightning and have never recovered* » [Swanson, 2001]. Il a réalisé que deux informations issues d'articles médicaux différents suggèrent lorsqu'on les juxtapose une hypothèse que personne ne connaissait alors. La connexion de deux informations disjointes peut créer une nouvelle information. Son approche était plus intuitive que structurée. En 1986, dans un article publié un an plus tard, il regrette de ne pouvoir décrire de processus systématique de recherche de connexions cachées [Swanson, 1987]. Mais il élabore

[1] American Society for Information Science and Technology. www.asis.org

[2] La maladie de Raynaud est caractérisée par un arrêt temporaire de la circulation du sang au niveau des extrémités. Les doigts deviennent pâles et très douloureux. La maladie est favorisée par le froid.

rapidement une stratégie basée sur l'utilisation de bases des données bibliographiques Medline®[3], Embase®[4] et SciSearch®[5], baptisée *explore/exclude* ou *trial-and-error*. Cette stratégie permet de rechercher les connections entre deux articles (*literatures*), non interactifs (ne se citent pas) et complémentaires afin de générer une nouvelle information absente des deux articles considérés séparément [Swanson, 1989]. Son travail portera principalement sur l'amélioration de sa méthode de KDD et la découverte de nouvelles hypothèses.

## 2 – AUTRES VALIDATIONS EXPERIMENTALES

Le modèle initial qui avait permis d'établir le rôle de l'huile de poisson dans la maladie de Raynaud a fait l'objet par l'auteur de plusieurs validations expérimentales.

La deuxième étude de Swanson porte sur les connections qui existent entre la migraine et le magnésium [Swanson, 1988]. Après avoir identifié 11 connections « négligées »[6], il formule l'hypothèse qu'une déficience en magnésium d'origine alimentaire pourrait être une cause de la migraine.
A travers la bibliographie, Swanson ne trouve que deux articles qui établissent clairement un lien entre migraine et magnésium.
L'importance du magnésium dans le développement de la migraine est aujourd'hui clairement établie même si son rôle précis reste inconnu [Mauskop, 1998].

Dans une autre étude, Swanson a essayé de mettre en lumière les liens possible entre l'arginine et l'I*nsulin-like Growth Factor I* (IGF I ou somatomédine C) [Swanson, 1990].

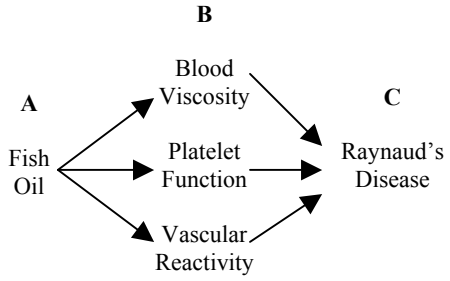
## 3 – MODELISATION DE LA DECOUVERTE

### 3.1 – Le modèle ABC

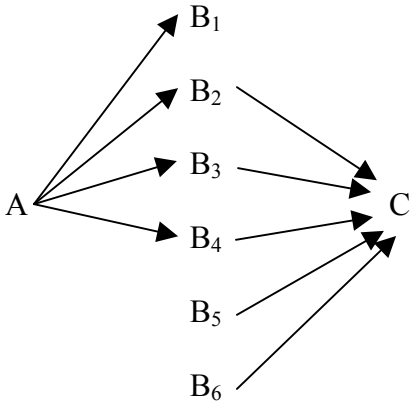
Swanson élabora le raisonnement suivant, soit :
- A l'huile de poisson,
- B l'agrégation plaquettaire, la réactivité vasculaire et la viscosité du sang,
- C la maladie de Raynaud.

A améliore C en agissant sur B. C'est le schéma classique d'action d'un médicament. Une maladie C est caractérisée par un certain nombre de désordres physiologiques B (physiopathologie), le médicament A agit favorablement sur les désordres physiologiques.

Dans ce premier exemple, Swanson connaissait les liens $A \rightarrow B$ et $B \rightarrow C$, et $A \cap C = \varnothing$. La physiologie représente l'élément commun qui permet de lier la maladie au traitement. B peut représenter plusieurs éléments. Dans le cas de la maladie de Raynaud, Swanson citera l'agrégation plaquettaire, la viscosité du sang et la réactivité vasculaire.




Plus A et C ont d'éléments B communs, plus il y a de chance que le lien AC soit fort et que l'on trouve par la suite des preuves expérimentales qui valideront l'hypothèse générée.




[3] Base de données bibliographique biomédicale produite par la National Library of Medicine (Bethesda, Mariland), accessible par l'interface PubMed : http://www.ncbi.nlm.nih.gov

[4] Base de données bibliographique biomédicale produite par Elsevier Science BV (Amsterdam), accessible par les serveurs DataStar™, STN® ou Dialog® (entre autres).

[5] Base de données bibliographique scientifique, dont la particularité est de permettre la recherche par référence citée. Elle est produite par Thomson ISI® et est accessible par les serveurs DataStar™, STN® ou Dialog® (entre autres).

[6] Dépression envahissante corticale (*spreading cortical depression*), épilepsie, substance P, agrégation plaquettaire, libération de sérotonine, bloqueur des canaux calciques, stress et profil comportemental de type A, tonus et réactivité vasculaire, prostacyclines et prostaglandines, inflammation, hypoxie.

### 3.2 – Le savoir public caché

Sachant que A agit sur B (sans en être la cause exclusive) et B agit sur C, on peut formuler l'hypothèse que A agit sur C. C'est un raisonnement par inférence ou transitivité. Si les liens AB et BC sont connus mais pas AC, Swanson parle de savoir public non-découvert (*undiscovered public knowledge*) ou caché. Nous considérons habituellement que les hypothèses sont inventées et non découvertes. Cependant dans le cas où AB et BC sont connus, alors l'hypothèse « A cause C » pré-existe implicitement, même si elle est inconnue, jusqu'à ce qu'on la découvre. Mettre AC en avant n'éclipse pas le fait qu'il s'agisse d'une hypothèse et que pour la valider, il faudra la confronter à l'expérimentation. L'existence de données décrivant les liens AB et AC la rendent plausible. A et C sont connectés de manière logique, mais bibliographiquement disjoint.

Swanson travaille principalement sur Medline®, accessible gratuitement, qui comprend aujourd'hui plus de 12 million de citations. 460 000 nouvelles citations ont été ajoutées en 2001. Faisons l'approximation suivante : si chaque article propose une relation de type AB ou BC, alors Medline® contient 72 000 milliard de combinaisons potentielles. Le savoir est caché par la masse d'information, l'expression *undiscovered public knowledge* prend ici tout son sens. L'information utile se trouve noyée dans la masse de données : il est ainsi pratiquement impossible à un chercheur de suivre les publications de sa propre discipline et strictement impossible de suivre systématiquement celles issues d'autres disciplines qui pourraient contribuer à l'avancement de ses travaux [Grivell, 2002].

### 3.3 – La méthodologie *explore/exclude* ou *trial-and-error*

S'appuyant sur l'exemple de la maladie de Raynaud/huile de poisson, Swanson proposera une méthodologie de KDD [Swanson, 1989]. Elle se décompose en deux parties de deux étapes. Il s'agit d'abord d'analyser la littérature, sur un sujet donné, pour identifier les connections logiques qui caractérisent ce sujet. C'est l'étape exploratoire qui fait appel à la créativité humaine. Puis la seconde partie a pour objectif d'exclure toutes les connections connues. Medline® est la base de données bibliographique utilisée.

I$^{ère}$ partie : exploration

*Etape 1 :* dans la littérature biomédicale, les titres des articles signalent souvent des concepts en relation avec le thème principal de l'article. Swanson met à profit cette particularité pour identifier la viscosité du sang et d'autres propriétés sanguines comme des facteurs sur lesquels agir pour traiter ou soulager les personnes atteintes de la maladir de Raynaud.

*Etape 2 :* une seconde recherche Medline® sur la viscosité du sang permet d'identifier les moyens de la modifier. Ainsi, l'huile de poisson apparaît comme diminuant la viscosité sanguine et agit favorablement sur les autres propriétés sanguines.

II$^{ème}$ partie : exclusion

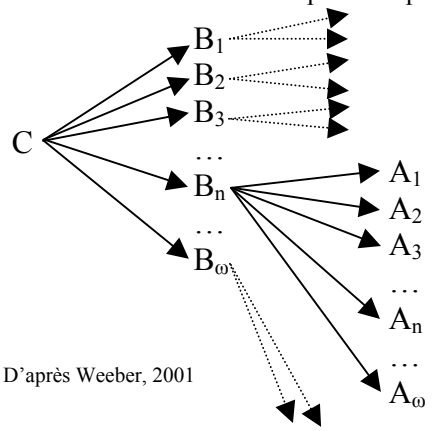
*Etape 3 :* l'hypothèse huile de poisson/maladie de Raynaud est-elle connue ? Une recherche est conduite pour vérifier qu'aucune référence bibliographique ne mentionne ensemble huile de poisson et maladie de Raynaud.

*Etape 4 :* déterminer si l'hypothèse de l'apport d'huile de poisson par l'alimentation pour traiter la maladie de Raynaud est médicalement recevable, en étudiant minutieusement les deux littératures.

Cette méthodologie, qui permet l'identification d'informations complémentaires et non-liées, n'est pas complètement automatisée et comprend une partie manuelle importante, sur laquelle Swanson travaillera par la suite. Il développera et automatisera sa méthode pour aboutir à la réalisation du projet Arrowsmith[7] (en collaboration avec Neil Smalheiser).

### 3.4 – Processus de découverte ouvert ou fermé

La *génération* d'hypothèses suit un processus ouvert. Le point de départ est la littérature C, connue, le but étant d'identifier B puis A, inconnus à priori. D'autres variations sont aussi possibles, comme A→B→C ou A←B→C par exemple.

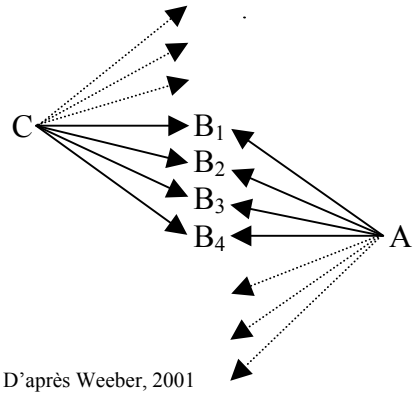



D'après Weeber, 2001

[7] Première version http://kiwi.uchicago.edu, puis http://arrowsmith2.psych.uic.edu.

Les flèches pleines représentent les voies intéressantes, les pointillés des échecs.
La stratégie *trial-and-error* est un processus ouvert : partant de la maladie de Raynaud, Swanson recherche les dérèglements physiologiques impliqués dans la maladie, puis identifie un traitement potentiel à priori inconnu [Swanson, 1989].

Un processus fermé *teste* une hypothèse, identifie les relations entre A et C : quels sont les liens entre la migraine et le magnésium [Swanson, 1988] (voir § 2) ? Il est aussi possible de tester d'autres combinaisons comme les liens entre A et B : les effets de l'arginine s'exercent-ils bien par l'intermédiaire de l'IGF I [Swanson, 1990] (voir § 2) ?



D'après Weeber, 2001

D'un point de vue général, tester une hypothèse revient à travailler sur un volume d'information plus restreint que de générer une hypothèse. Des outils de NLP (*Natural Language Processing*) et le recours aux experts permettent de réduire le nombre de documents à traiter manuellement.

## 4 – CONCLUSION

Le mythe du « savant homme » du 19ème a laissé la place, avec le développement des techniques, à une compartimentation des spécialités et à un accroissement formidable de la littérature disponible. Swanson explique que cela n'est pas sans poser de problèmes, qui méritent que nous les examinions. Tout d'abord, la plupart des disciplines scientifiques sont certainement reliées à d'autres de manière logique. Ensuite, il existe bien plus de combinaisons possibles entre discipline scientifiques qu'il y a de disciplines. Enfin, le système de structuration de l'information dans les bases de données bibliographiques n'est pas organisé pour exploiter et valoriser les connexions, beaucoup nous échappent. Ainsi sont créées un grand nombre d'unités de littérature, indépendamment les unes des autres, sans tenir compte des relations logiques qui peuvent les lier et donner naissance à de nouveaux savoirs : il s'agit du savoir public caché [Swanson, 1986]. La science répond à sa propre croissance par une augmentation de la spécialisation en négligeant les connexions.

Le modèle de Swanson tire partie de cette organisation fragmentée et de plus en plus cloisonnée de l'information scientifique et de la quantité colossale de publications disponibles : en combinant des informations bibliographiques existantes, bien que non reliées, on peut créer de nouvelles connaissances. Les points importants du modèle sont l'absence de lien entre les deux informations et leur complémentarité : cela conditionne l'existence d'une relation cachée.

## BIBLIOGRAPHIE


Dewailly, E., Blanchet, C., Lemieux, S., Sauvé, L., Gingras, S., Ayotte, P ., Holub, B.J. (2001), " *n-3 Fatty acids and cardiovascular disease risk factors among the Inuit of Nunavik* ", American Journal of Clinical Nutrition. Vol. 74, n°4, p. 948-954.

DiGiacomo, R.A., Kremer, J.M., Shah, D.M. (1989), "*Fish-oil dietary supplementation in patiens with Raynaud's phenomenon : a double-blind, controlled, prospective study* ", American Journal of Medicine. Vol. 86, n°2, p. 158-164.

Grivell, L. (2002), "*Mining the bibliome : searching for a needle in a haystack ?* ", EMBO Reports. Vol. 3, n°3, p. 200-203.

Mauskop, A., Altura, B.M. (1998), "*Role of magnesium in the pathogenesis and treatment of migraines* ",Clinical Neuroscience. Vol. 5, n°1, p. 24-27.

Swanson, D.R. (1986), "*Fish oil, Raynaud's syndrome, and undiscovered public knowledge* ", Perspectives in Biology and Medicine. Vol. 30, n°1, p. 7-18.

Swanson, D.R. (1987), "*Two medical literatures that are logically but not bibliographically connected* ", Journal of the American Society for Information Science. Vol. 38, n°4, p. 228-233.

Swanson, D.R., (1988), "*Migraine and magnesium : eleven neglected connections* ", Perspectives in Biology and Medicine. Vol. 31, n°4, p. 526-557.

Swanson, D.R. (1989), "*Online search for logically-related noninteractive medical literatures : a systematic trial-and-error*

*strategy* ", Journal of the American Society for Information Science. Vol. 40, n°5, p. 356-358.

Swanson, D.R. (1990), "*Somatomedin C and arginin : implicit connections between mutually-isolated literatures* Perspectives in Biology and Medicine. Vol. 33, n°2, p. 157-186.

Swanson, D.R. (2001), "*ASIST Award of Merit acceptance speech : on fragmentation of knowledge, the connection explosion, and assembling other people's ideas*", Bulletin of the American Society for Information Science and Technology. Vol. 27, n°3, www.asis.org.

Weeber, M., Klein, H., de Jong-van den Berg, L.T.W, Vos, R. (2001), "*Using concepts in literature-based discovery : simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries*", Journal of the American Society for Information Science and Technology. Vol. 52, n°7, p. 548-557.