Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação
E-ISSN: 1518-2924
bibli@ced.ufsc.br
Universidade Federal de Santa Catarina
Brasil

Polanco, Xavier
Transformer L'information en connaissance avec stanalyst. Cadre conceptuel et modele
Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, núm. Esp, primer semestre, 2008, pp. 76-91
Universidade Federal de Santa Catarina
Florianopolis, Brasil

TRANSFORMER L'INFORMATION EN CONNAISSANCE AVEC STANALYST. CADRE CONCEPTUEL ET MODELE.

TURNING INFORMATION INTO KNOWLEDGE WITH STANALYST. CONCEPTUAL FRAMEWORK AND MODEL.

Xavier Polanco - xavier.polanco@lip6.fr / xavier.polanco@inist.fr
Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie – Paris 6
Institut de l'Information Scientifique et Technique
Centre National de la Recherche Scientifique

Résumé
STANALYST : A quoi sert-il ? Comment est-il fait ? A quelle conception réponde-t-il ? Nous allons essayer de donner réponse à ces trois questions nous situant dans le contexte de la science de l'information. Ici nous nous concentrons essentiellement sur la fonction et la conception qui sont à la base de STANALYST en laissant de côté son mode d'emploi. Ce n'est pas un discours technologique au sens de rester exclusivement au niveau de la description d'une technologie de l'information. Au contraire, la tâche que nous nous imposons est de placer la technologie dans le cadre d'une certaine conception du travail sur l'information. Notre sujet central est démontrer qu'est-ce que l'analyse de l'information et comment nous pouvons réaliser cette analyse avec un outil conçu pour produire une information élaborée, et plus encore pour transformer l'information en connaissances. L'intérêt du sujet est encore plus sensible à l'heure actuelle où la société est reconnue comme société de l'information, et l'on estime en outre que celle-ci évolue vers une société de la connaissance.

Mots clef: Technologie de l'intelligence. Analyse de l'Information. Information Élaborée. Bibliométrie. Infométrie

1 INTRODUCTION

Dans son sens le plus général, STANALYST est une technologie de l'information spécialisée pour l'analyse de l'information. STANALYST est le produit de recherches menées dans l'INIST / CNRS. Elle est accessible depuis un navigateur web, et intègre sous une interface commune un ensemble de programmes informatiques. Pour des raisons de circonstance, mais aussi de droit de propriété, STANALYST se trouve actuellement lié aux seules bases PASCAL et FRANCIS de l'INIST-CNRS[1] plus les bases SciELO[2].

[1] http://www.inist.fr/
[2] http://www.scielo.org/

La compatibilité de STANALYST avec les bases SciELO est le résultat d'un projet de coopération multilatéral entre organismes de la France, l'Argentine, le Brésil et le Chili, soutenu par le Ministère des Affaires Etrangères de France, entre 2005-2006, à l'initiative de la Délégation régional de coopération France-Cône Sud[3]. Dans ce projet participèrent l'Institut de l'Information Scientifique et Technique du Centre National de la Recherche Scientifique (INIST-CNRS), le Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME / OPS / OMS), le Centro Argentino de Información Científica y Tecnológica (CAICYT), la Comisión Nacional de Investigación Científica y Tecnológica du Chili (CONICYT), et la Red Iberoamericana / Interamericana de Indicadores de Ciencia y Tecnología (RICYT)[4].

La première référence à STANALYST est de 2001 où cette station d'analyse est présentée comme « an integrated environment for clustering and mapping analysis on science and technology » en soulignant le « linguistic processing module » et les « clustering programms » (Polanco et al, 2001) . C'est seulement dans un poster en 2004 que STANALYST est pour la première fois décrit tel qu'il existe aujourd'hui (Besagni et al, 2004). La version compatible avec les bases SciELO fut elle exposée en janvier 2006 dans un séminaire international, à Santiago du Chili (Polanco 2006) ; et puis, en novembre de la même année, à Buenos Aires dans les journées internationales d'étude comparative France-Amérique Latine sur « el espacio público de las ciencias sociales y humanas », publiées en 2007 dans un livre (Polanco 2007). Dans ce contexte de chercheurs en sciences humaines et sociales, fut spécialement souligné l'aide que STANALYST pouvait signifier au travail académique et à la recherche. Quelque temps après, à la conférence internationale de RICYT, mai 2007, à São Paulo, fut soumise une contribution où la présentation de STANALYST-SciELO est suivie d'une application sur des données sur le cancer extraites des bases SciELO, Argentine, Brésil et Chili. Cette application illustrait le type d'analyse que l'on peut faire avec STANALYST à partir des bases SciELO. Finissons cette synopsis en rappelant que sur le plan opérationnel, STANALYST est encore dans une phase expérimentale[5].

La suite de l'article s'organise de la manière suivante : la section 2 répond ce à quoi sert STANALYST mettant en valeur la notion d'analyse de l'information et de ce qu'elle représente. Qu'est-ce qu'on peut faire avec STANALYST ? La section 3 détaille les aspects de technologie de l'information et plus encore de technologie de l'intelligence, insistant alors sur la structure et les techniques mises en œuvre. La section 4 développe la conception épistémologique à la base de STANALYST et à laquelle cet outil représente une réponse technologique. Comme il est traditionnel, une conclusion close l'article.

[3] http://www.france-conesud.cl/
[4] BIREME http://www.bireme.br/ ; CAICYT http://www.caicyt.gov.ar/ ; CONICYT http://www.conicyt.cl/ ; RICYT http://www.ricyt.edu.ar/
[5] En phase beta test, STANALYST-bases FRANCIS et PASCAL est accessible à l'INIST http://stanalyst.inist.fr/ ; en phase alpha test, STANALYST-SciELO est accessible à BIREME http://turquesa.bireme.br:8080/ et au CAICYT http://lan.caicyt.gov.ar:3000/

2 PREMIERE QUESTION : A QUOI SERT-IL ?

STANALYST sert pour l'analyse de l'information. Mais alors que devons-nous entendre par analyse de l'information au-delà de sa compréhension intuitive ? Cette section est un essai de réponse.

2.1 Traitement de l'information

On distingue trois phases dans le traitement de l'information : d'abord le stockage, puis l'accès ou la recherche d'information et enfin l'analyse elle-même. D'un point de vue logique, on peut considérer ces trois phases comme les fonctions constitutives d'un système d'information : le stockage et gestion des données, les mécanismes de recherche d'information et puis l'analyse de l'information. En effet, une fois que l'information qui nous intéresse est disponible dans des bases de données ou bien dans le web, et nous disposons de systèmes de recherche d'information ou de moteurs de recherche, le problème est alors celui d'analyser l'information collectée. Le présupposé est que la quantité d'information collectée dépasse les moyens humains ordinaires et qu'il faut donc s'aider de moyens informatiques adéquats. Autrement dit, le stockage, la recherche et la diffusion d'information sont des activités caractéristiques du traitement de l'information auxquelles vient s'ajouter l'analyse de l'information. Ce qui importe c'est de distinguer l'objectif et la forme de l'analyse de l'information. Les objectifs dépendent des secteurs d'activité où l'information représente une valeur qu'il est d'intérêt exploiter, une matière à transformer en information utile, à transformer en connaissances. Quant à la forme de le faire, il y a différentes modes d'implémentation de ce travail de transformation.

2.2 Analyse de l'information

Commençons par la définition que nous avons adoptée (Polanco 1997) comme guide de nos travaux conduisant à STANALYST. Pour ensuite, revenir sur la place de l'analyse dans un système d'information.

Notre définition correspond à l'analyse assistée par ordinateur. En général, par l'analyse de l'information, on entend la phase d'interprétation que l'utilisateur réalise d'une manière directe et manuelle. Les limites de ce type d'analyse sont évidentes du moment où il s'agit de traiter une quantité importante de données, cela d'une part, et d'autre lorsque nous souhaitons incorporer l'analyse dans un système de production d'information élaborée. Nous avons appelé analyse de l'information l'application (1) de techniques statistiques (c'est-à-dire bibliométriques), (2) de traitement automatique du langage naturel, (3) de classification automatique non supervisée et (4) de représentation graphique (cartographie) du contenu cognitif et factuel des données bibliographiques. Cette définition est opérationnelle au moyen des technologies que nous avons conçu et développé dont STANALYST est l'aboutissement.

L'objectif commun de ces quatre processus est de signaler les centres d'intérêt, les thèmes ou les topiques contenus dans une quantité d'information disponible en langage naturelle (données textuelles), et autour desquels s'agrègent les divers éléments de cette information (i.e. articles, périodiques, auteurs, laboratoires, pays). Ensuite, de visualiser

les thèmes ou topiques sur une carte afin d'apprécier leurs positions relatives.

Maintenant si nous cherchons à placer l'analyse telle quelle vient d'être définie dans un système d'information, nous avons besoin de contraster l'analyse de l'information avec le stockage de données et la recherche d'information, mais aussi avec l'extraction d'information et l'acquisition de connaissances à partir des bases de données. En faisant cela nous contribuons à définir la place de STANALYST dans un système d'information.

Au sens le plus large du terme : « Information retrieval deals with the representation, storage, organization of, and access to information items » (Baeza-Yates & Ribeiro-Neto, 1999). Comme le rappellent les auteurs eux-mêmes, ils ont repris la distinction entre « recherche de données » (*data retrieval*) et « recherche d'information » (*information retrieval*) de van Rijsbergen (1979). Une distinction que nous pouvons étendre à l'analyse de l'information pour distinguer celle-ci de l'analyse de données, et souligner que la relation entre information et donnée n'est pas une relation d'identité. L'analyse de données fait partie de l'arsenal de techniques statistiques que l'analyse de l'information peut mettre en œuvre pour atteint ses fines, c'est donc un moyen pour que l'analyse de l'information réalise ses propres objectifs.

Dans le contexte de la recherche d'information (c'est-à-dire, de l'*information retrieval*), la recherche de données (*data retrieval*) consiste principalement en déterminer quels documents dans une collection contiennent les mots clés de la requête que l'utilisateur a formulé. Comme Baeza-Yates et Ribeiro-Neto observent, l'utilisateur d'un système de recherche d'information est plus concerné par la recherche des informations sur un sujet d'intérêt que par la recherche des données satisfaisant une requête. Un langage d'extraction de données, comme SQL par exemple, vise essentiellement à rendre opérationnelle l'interrogation d'une base de données relationnelle, en revanche la recherche d'information a trait au langage naturel dans l'objectif de satisfaire le besoin d'information de l'utilisateur, de ce fait la notion de « pertinence » est au centre de la recherche d'information.

A ces deux paliers, « *data retrieval* » et « *information retrieval* » dans le traitement de l'information, vient s'ajouter un troisième où l'objectif est convertir l'information pertinente en connaissance à propos d'un sujet déterminé. L'analyse suppose comme nous l'avons dit que les systèmes de stockage de données et de recherche d'information existent et sont efficaces. A partir de là, le problème de l'analyse de l'information est celui d'exploiter l'information déjà trouvée, autrement dit de la traiter en tant qu'objet d'étude, visant en dernière instance l'acquisition et la représentation de connaissances. La formule est donc un cycle comme le suggère la figure 1.

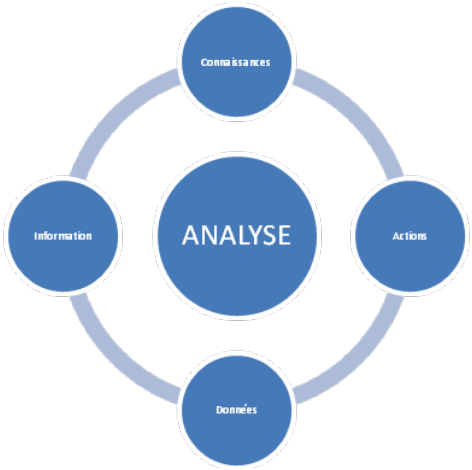

Figure 1 : Le cycle de l'analyse de l'information

L'analyse de l'information est proposée comme le système permettant d'extraire de l'information utile à partir des données, puis de transformer l'information en connaissances qui deviennent elles par la suite des actions qui génèrent à nouveau des données et le cycle recommence (cette idée nous l'avons esquissée dans Polanco 1999). L'analyse de l'information est donc une fonction spécifique à l'égard des autres deux fonctions (stockage et recherche d'information). L'analyse de l'information constitue une fonction propre nécessitant une recherche spécifique et des implémentations technologiques particulières. Ceci ne signifie pas qu'elle ne s'intègre pas dans un système d'information global réunissant les trois fonctions.

L'analyse de l'information est une expression générique qui présente des rapports générique-spécifique avec le domaine que l'on connaît comme la « découverte de connaissance dans les bases de données » (en anglais *knowledge discovery in databases*) (Fayyad et al, 1996 ; Piatetsky-Shapiro & Frawley 1991), et la « fouille de données » (*data mining*) (Maimon & Rokach 2005) ; si les données sont des textes, alors on parle de « fouille de textes » (*text mining*). Dans le cas qui nous intéresse ici, l'analyse se traduit dans la fouille de données textuelles pour l'acquisition et représentation de connaissances. En effet, le concept d'analyse de l'information apparaît comme le dénominateur commun des toutes ces opérations où l'information représente une matière première qu'il faut traiter afin d'obtenir une information utile. Rappelons au passage que depuis ses origines la découverte ou extraction de connaissance est définie presque toujours dans ces mêmes termes : « the non-trivial extraction of implicit, unknown, and potentially useful information from data » (Frawley et al, 1991).

A côté de la recherche d'information champ traditionnel et central de la science de l'information (van Rijsbergen 1979 ; Salton & McGill 1983), à partir de 1987, s'est développé un nouveau champ sous le nom d'extraction d'information (*information extraction*). Les *Message Understanding Conferences* (MUC) ont été lancées et financées par le DARPA pour encourager l'élaboration de nouvelles et meilleures

méthodes d'extraction d'information. On peut lire dans le site officiel de cette initiative : « Information Extraction is a technology that is futuristic from the user's point of view in the current information-driven world. Rather than indicating which documents need to be read by a user, it extracts pieces of information that are salient to the user's needs. Links between the extracted information and the original documents are maintained to allow the user to reference context »[6].

L'extraction d'information est donc une technologie de l'information qui se distinguerait de la recherche d'information par l'extraction de morceaux d'information qui sont saillantes aux besoins de l'utilisateur. En général, dans le contexte du traitement de l'information, technologie et science se trouvent intimement entrelacées. Souvent la technologie de l'information soulève des problèmes qui provoquent de nouvelles recherches au niveau de la science de la computation. Cette remarque concerne aussi STANALYST.

2.3 Etude de la science et de la technologie

Afin d'éclairer la position de l'analyse de l'information dans le contexte des études de la science et de la technologie, nous utilisons le schéma de Leydesdorff (1989) sur l'existence de trois dimensions et que nous formulons ainsi : les chercheurs (scientifiques et ingénieurs) travaillant dans des organisations produisent des connaissances qu'ils communiquent sous la forme de textes (articles, brevets, notes, documentation technique), chacune de ces dimensions (les chercheurs, les documents et les connaissances) correspond à un objet d'analyse ou d'étude pour la sociologie des sciences et des techniques, pour la science de l'information et pour les approches comportant le suffixe « métrie » ou « métrique ».

Tel que le montre la figure 2, nous proposons de revoir le schéma tridimensionnel sous la forme d'un cycle. Le cycle met en évidence le flux entrées et sorties dans le processus de production de connaissances scientifiques et technologiques.

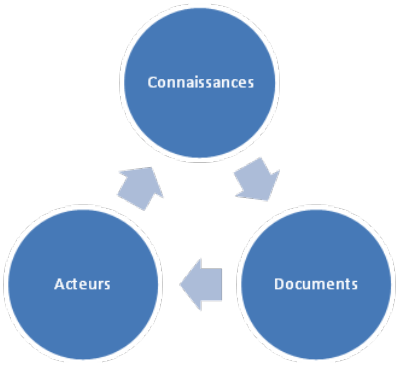

Figure 2 : Le cycle du schéma tridimensionnel de Leydesdorff (1989).

[6] Citation 25/09/2007 : http://www.itl.nist.gov/iad/894.02/related_projects/muc/index.html

Ce schéma permet d'observer qu'il y a une différence entre documents et connaissances ; comme l'avait déjà remarqué Brooks (1980), « document and knowledge are not identical entities ». Ainsi, la métrique appliquée sur les documents (bibliométrie) ne traduit pas directement une mesure des connaissances contenues dans les documents. Mesurer des documents (untel a produit *n* articles) ne signifie pas mesurer des connaissances car le nombre d'articles n'est pas égal au nombre de connaissances produites, articles ≠ connaissances. D'autre part, le schéma permet également d'observer que la connaissance dans l'espace mentale des chercheurs et ingénieurs peut se distinguer de la connaissance que l'on extrait des textes (de leurs publications). Les connaissances existant dans l'espace mental d'un individu font de lui un expert dans un domaine donné. Et son étude relève de la psychologie cognitive. La connaissance contenue dans les publications constitue une connaissance objective (au sens de Popper 1972). Les connaissances ont deux modes d'exister : l'un dans l'espace mental de leurs producteurs et l'autre dans des documents. Cette dernière forme est l'objet de notre approche, tout en gardant à l'esprit la remarque de Brooks (1980).

Il est intéressant de noter que, dans le champ de la scientométrie, la méthode de mots associés (*co-word analysis*) a été proposée depuis le début comme une « scientométrie cognitive » (Rip & Courtial, 1984) ou « scientométrie qualitative » (Callon et al, 1986), car son objectif via les associations de mots clés (cooccurrences) est de saisir le contenu cognitif des documents, ce que l'analyse de citations et co-citations ne fait qu'indirectement (Callon et al, 1993). STANALYST s'inscrit justement dans cette tradition inaugurée par le programme LEXIMAPPE au début des années quatre-vingts (Callon et al, 1983).

Les mots que les acteurs emploient, ainsi que les mots clés indexant les documents, envoient vers de concepts et les concepts à leur tour renvoient à des objets métalinguistiques dans le monde (objets physiques ou abstraits). Ceci est connu en linguistique sémantique comme le triangle sémiotique. S'agissant de documents, c'est le langage écrit qui importe ici de traiter (vis-à-vis du langage parlé ou parole). Le module INDEXATION de STANALYST doit donc être revue de cette perspective : c'est au travers des mots ou de termes et leurs variations que l'on accède aux concepts et connaissances. Ce module réalise pour le moment seulement une analyse morphologique et syntaxique du langage naturel manquant d'une couche d'analyse sémantique.

3 DEUXIEME QUESTION : COMMENT EST-IL FAIT ?

3.1 Technologie de l'intelligence

STANALYST est d'abord une technologie conçue pour l'analyse de l'information et dans cette mesure nous pouvons dire aussi que STANALYST est une technologie de l'intelligence. L'expression technologie de l'intelligence a été proposée par Levy (1990) et c'est donc de lui que nous l'empruntons pour désigner les outils informatiques d'aide à l'analyse. D'autre part, elle nous a été suggérée par l'intelligence économique et la veille technologique et scientifique. Mais si nous considérons l'intelligence artificielle,

alors nous voyons que celle-ci cherche à introduire de l'intelligence dans les systèmes d'information, particulièrement dans les systèmes à base de connaissances où l'apprentissage et le raisonnement sont mis en place. Mais ici nous devons distinguer entre d'une part produire des « technologies intelligentes » et d'autre part produire des « technologies de l'intelligence ». Pourtant, il est permis de parler de « technologies intelligentes de l'intelligence ». Ce à quoi nous devons nous orienter car STANALYST n'est pas encore à ce stade. Le projet est ouvert.

3.2 Modèle

Modéliser la notion d'analyse de l'information s'impose afin de pouvoir le donner une implémentation opérationnelle, c'est-à-dire technologique. Comme nous l'avons souligné (Polanco 2007a), STANALYST constitue une technologie de l'information au service du travail intellectuel, comparable à un traitement de texte ou à un tableur, auxquels nous sommes habitués à nous en servir pour nos travaux dans l'enseignement et la recherche. L'idée de base est que l'analyse suppose comme préalable en avoir un projet en tête à propos de ce que l'on veut analyser. Ensuite, la démarche consiste à délimiter un domaine d'étude sous la forme d'un corpus sur lequel se suivent des opérations de type quantitatif, qualitatif et de classification. La traduction technologique de ces opérations intellectuelles est réalisée par des modules qui produisent un certain nombre de résultats et avec lesquels l'analyste ou le chercheur réalise sont propre travail intellectuel d'explication et d'interprétation. La figure 3 rendre visible d'une manière épurée ce que nous venons de dire.

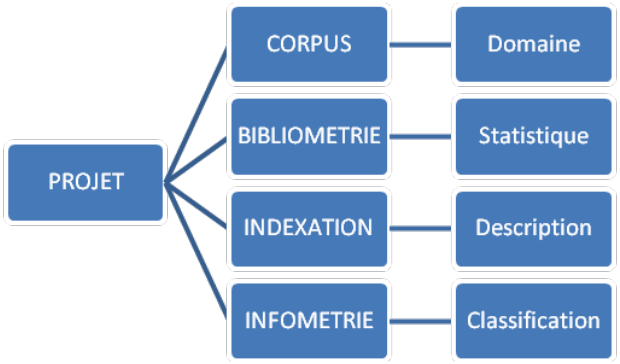

Figure 3 : Le modèle STANALYST. Suite à la définition d'un projet apparaissent les modules opérationnels accompagnés ici par les mots clés génériques de la connaissance qu'ils produisent respectivement.

L'analyse comprend d'abord une phase statistique dont l'objectif est de connaître les fréquences et distributions des données bibliographiques et de leurs composants (comme par exemple auteurs, pays d'affiliation, revues et leurs pays d'édition, mots clés, types de documents et date de publication) ; cette phase statistique connue comme bibliométrie est suivie d'une phase où l'aspect linguistique ou textuel des données est primordial, en bref, titres, résumés et indexation. Titres et résumés sont à la source de

l'indexation automatique. On sait que les mots clés expriment les concepts qui sont présents dans les documents. C'est à partir de la matrice D = (N, P), N documents × P mots clés, que l'on passe à la phase de classification automatique dont le produit est l'organisation des documents et des mots clés en classes, et les classes sont visualisables sur des cartes. L'analyste actionnant les modules du système obtient des résultats, c'est-à-dire de l'information à propos des données collectées (corpus), qu'il peut par la suite utiliser pour faire un travail d'interprétation ou d'analyse plus approfondie. On remarquera que nous ne sommes pas encore avec STANALYST au niveau du texte plein (*full-text*), les données sont toujours des références bibliographiques extraites des bases de données et donc des données structurées.

3.3 Architecture et fonctionnement

L'architecture de STANALYST est la suivante. L'ACCUEIL est une page HTML statique où l'utilisateur déclare son nom et son mot de passe pour accéder aux différents modules. L'utilisateur crée alors un PROJET, c'est-à-dire un répertoire dans lequel seront stockés tous les résultats le concernant et définissant ainsi un environnement de travail. Il en est le propriétaire, mais il a également la possibilité de donner accès à son projet aux utilisateurs associés reconnus par la station. Les modules CORPUS, BIBLIOMÉTRIE, INDEXATION et INFOMÉTRIE constituent les modules de travail. Ces modules exécutent des programmes qui opèrent sous la commande de l'utilisateur.

Le module CORPUS permet de rédiger et exécuter une requête, ce qui a pour effet de générer un corpus qui sera utilisé par les modules suivants. Le module BIBLIOMÉTRIE supporte la production de statistiques descriptives qui fournissent une information quantitative et que l'on peut ensuite traiter conformément aux lois bibliométriques (Bradford, Lotka, Zipf). Le module INDEXATION permet de réviser l'indexation préexistante en vue de la classification thématique. Il autorise également de procéder à une indexation automatique du corpus, en s'appuyant pour cela sur ILC, un ensemble d'outils d'ingénierie linguistique réalisant une indexation contrôlée à partir de plusieurs référentiels terminologiques (comme il est exposé notamment dans Royauté 1999 ; Daille et al, 2001 ; Jacquemin et al, 2002). Le module INFOMÉTRIE permet une classification thématique à l'aide de deux programmes de classification automatique non supervisée, NEURODOC et SDOC (Grivel & François 1995), le premier exécute une classification non hiérarchique basée sur une variante de k-means, appelée k-means-axial (Lelu 1993) ; le second une classification hiérarchique basée sur la cooccurrence de mots. L'emploi de ces méthodes de classification automatique sert à détecter des thèmes ou des centres d'intérêt à partir des données. L'utilisateur dispose ainsi de deux points de vue qu'il peut comparer grâce à une fonction du module INFOMETRIE. Rappelons au passage que les méthodes de classification automatique relèvent de l'analyse de données (Saporta 1990 ; Lebart et al, 1997). La visualisation des classes sur une carte offre un moyen d'évaluer la position des thèmes (classes) sur un plan de représentation. Les cartes constituent un des moyens de la visualisation de l'information (Card et al, 1999). Le rôle de la visualisation est primordial dans l'analyse de

l'information (Brachman et Anand 1996) [7].

Les modules présentent tous une même organisation graphique composée de trois fenêtres : la première affiche l'historique du processus, la seconde correspond au lancement des opérations propres à chaque module, et la troisième permet la gestion des résultats. Pour la mise en œuvre des opérations, l'utilisateur n'a qu'à cliquer sur les commandes signalées dans chacune des fenêtres. En entrée de chaque module l'utilisateur est sollicité de remplir un formulaire de paramètres déterminant selon les modules soit les statistiques souhaitées, soit le type d'indexation (automatique ou pas), soit la méthode et les variables de la classification automatique.

Actuellement, l'indexation automatique du module INDEXATION opère sur l'anglais et le français. De sorte que l'indexation automatique des données SciELO se fait exclusivement à partir d'articles dont les titres et les résumés sont en anglais. L'ambition serait de pouvoir élargir le traitement automatique de langues à l'espagnol et au portugais. Or, l'indexation automatique à partir de l'espagnol et du portugais reste conditionnée au développement d'outils linguistiques adaptés au traitement automatique de ces deux langues.

D'autre part, le fait de pouvoir travailler sur plusieurs langues pose le problème complexe du multilinguisme, c'est-à-dire d'être capable de traiter automatique au même temps des données de langues différentes et non pas séparément. Nous savons que le respect de la diversité linguistique dans le monde numérique dépend de plus en plus de la mise en place des solutions multilingues. Si nous considérons que STANALYST se trouve maintenant dans un contexte multilingue, la recommandation de l'UNESCO sur la promotion et l'usage du multilinguisme apparaît comme incontournable. Cependant, il ne faut pas se cacher la complexité technologique que la gestion multilingue implique dans un système d'information.

[7] L'actualité et l'importance de la visualisation se laisse sentir pour le nombre d'articles de synthèse publiés par l'*Annual Review of Information Science and Technology* entre 1995 et 2005 ; voir les volumes 39 (2005), 37 (2003), 32 (1997) et 30 (1995).

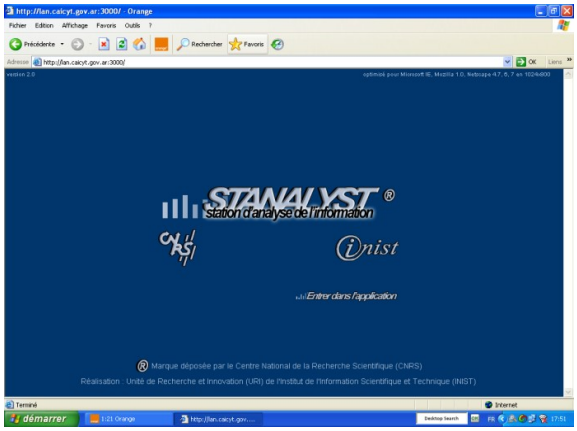

Figure 4 : Une prise d'écran de la page d'accès, telle qu'elle s'affiche aujourd'hui.

4 TROISIEME QUESTION : A QUELLE CONCEPTION REPONDE-T--IL ?

Nous nous réfèrerons à la conception de la connaissance qui sous-tend STANALYST en tant que technologie de l'intelligence. D'une part, la connaissance est comprise comme un processus de production (Althusser 1965) et, d'autre part, la connaissance qui nous intéresse ici est la connaissance objective (Popper 1972) comme il a été déjà évoqué dans la section 2. Cette double vision nous mène à considérer l'analyse de l'information comme un processus de production dont la matière première est la connaissance objective et son instrument de travail STANALYST. Le modèle tient comme base d'origine les théories philosophiques des « trois généralités » d'Althusser, et des « trois mondes » de Popper. Par ailleurs, ces deux formulations se ressemblent fortement.

Nous sommes d'accord avec la remarque de Brooks (1980) que le rôle de la science de l'information est l'exploration et l'organisation du troisième monde de Popper, celui de la connaissance objective. Dans sa contribution « The foundations of information science » (1980), après un rappel de la théorie poppérienne des trois mondes et de la connaissance objective, Brooks soutient : « What information science needs at its roots, it seem to me, is an objective rather a subjective theory of knowledge » (Part I, p. 127). Nous partageons cet avis depuis le commencement de nos travaux conduisant à la conception de STANALYST. Au même temps, nous avons été orientés par la théorie de la connaissance d'Althusser (1965) des trois généralités.

D'après Popper, il existe le monde de phénomènes physiques et sociaux (monde 1), le monde subjectif des états de conscience, des états mentaux et des dispositions behavioristes, celui du sujet connaissant (monde 2), et par rapport auquel la connaissance écrite, celle qui est véhiculée par la littérature scientifique que STANALYST permet d'analyser représente la connaissance objective (monde 3). Ceci

induit la reconnaissance qu'il y a deux catégories de problèmes concernant l'analyse de la connaissance. La première catégorie comprend les problèmes relatifs aux actes de génération ou de formation de connaissances ; la seconde comprend les problèmes ayant trait aux structures de la connaissance produite au sens d'écrite et publique. C'est la deuxième catégorie de problèmes qui constitue l'objet de notre travail. Il s'agit d'analyser l'état de la connaissance produite et couchée dans des documents écrits, afin de fournir une représentation de sa structure à un moment donnée de son développement. On ne cherche pas à capter la connaissance des sujets comme c'est le cas dans la tradition des systèmes experts. A notre avis, l'analyste ne doit pas s'occuper de la connaissance en action dans les compétences des individus (sujets de la connaissance), mais de la connaissance produite par eux et stockée dans les bases de données, l'objectif étant l'extraction de connaissances utiles pour la prise de décision, la définition de stratégies et l'évaluation de l'état de la science et de la technologie à un moment donné. Aujourd'hui, cette activité est représentée par la « découverte de connaissances dans les bases de données » comme il a été dit dans la section 2.

Dans le schéma d'Althusser (1965) des trois généralités, la généralité I (GI) designe les objets d'une science, la généralité II (GII) les moyens de travail théorique et la généralité III (GIII) est la connaissance que l'on produit comme résultat du travail de GII sur GI. En appliquant ce schéma à notre cas, les données seraient la GI que l'analyste transforme en connaissance GIII, au moyen de l'application d'une technologie de l'information telle que STANALYST, GII. On voit que le schéma correspond bien à la notion de processus de production, et auquel s'ajuste l'idée de la transformation des données en information et de celle-ci en connaissance ; il y a toujours d'un côté l'entrée et de l'autre la sortie du système opérant la transformation ou conversion ; c'est aussi l'idée de produire une information élaborée à partir d'une information brute. L'état de l'information, élaborée ou de brute, n'est pas une donne absolue ni nécessairement fixé par la nature des données, ces états de l'information dépendent à la fois du système mis en œuvre (comme par exemple STANALYST) et de son but, c'est-à-dire le besoin informationnel qu'il entend satisfaire.

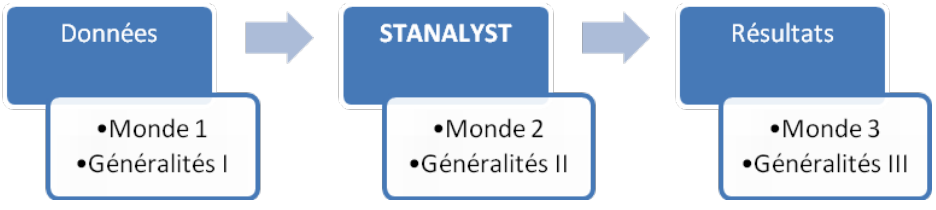

Figure 5 : Correspondance de STANALYST avec les « trois mondes » de Popper et les « trois généralités » d'Althusser.

La figure 5 montre STANALYST à la place de l'agent humain auquel font références le

monde 2 de Popper (M 2) comme les généralités II d'Althusser (G II) en tant que technologie opérant sur des données et produisant une nouvelle intelligence à leur sujet. C'est vrai que ces deux philosophes ont émis leurs théories avant l'ère de la société de l'information, elles relèvent plutôt de l'époque de la société industrielle. En revanche, nous réfléchissons et agissons de plein pied dans l'information et la gestion de connaissances, et nous nous en servons habituellement des technologies de l'information dans notre travail de producteurs de connaissances.

5 CONCLUSION

STANALYST : A quoi sert-il ? Comment est-il fait ? A quelle conception réponde-t-il ? Nous avons tenté de fournir les réponses à ces trois questions en mettant en valeur la notion d'analyse de l'information et ce qu'elle représente. Plus qu'une technologie de l'information c'est une technologie de l'intelligence, nous avons dit, insistant au même temps sur la structure et les techniques mises en œuvre par la station d'analyse. Et pour finir nous avons révélé son cadre épistémologique où s'assise l'analyse en tant que processus de transformation « données → information → connaissance ». Nous avons pu signaler le rôle d'agent technologique (artificiel) qu'un tel instrument joue dans le processus de production de connaissances, ou de traitement de la connaissance objective. Cela nous fait penser à l'univers des sciences de l'artificielle (Simon) où le sujet abordé viendrait finalement se positionner.

Pas question dans cet article de mode d'emploi ni de manuel d'utilisation, nous nous sommes attelé surtout à rester à un niveau conceptuel et de tour d'horizon. Le défaut de ce type d'approche est qu'elle laisse sous silence des aspects techniques et scientifiques importantes. La consolation est de penser que nous avons fourni au moins un cadre où placer la technologie que STANALYST représente et des moyens pour la mettre en perspective et si nécessaire la dépasser.

Et pour finir, un aveu, aujourd'hui l'auteur revoit STANALYST de ce quadruple point de vue : (1) des apports de l'apprentissage automatique (c'est-à-dire *machine learning*) dans le domaine de la classification, (2) de l'état d'avancement de la représentation des connaissances et du raisonnement (voir par exemple ce qui se fait à propos du web sémantique), (3) ainsi que de la sémantique dans l'ingénierie de langues, et (4) des apports de la théorie de graphes, hypergraphes et treillis à l'analyse de réseaux de connaissance. En somme, STANALYST est pour nous la source d'un nouveau programme scientifique. Autrement dit, le monde est ouvert et non pas clos.

REFERENCES

ALTHUSSER, L. **Pour Marx**. Paris: Maspero, 1965.
BAEZA-YATES, R., RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, Addison-Wesley, 1999..

BESAGNI, D., FRANÇOIS, C., POLANCO, X., ROCHE, I. Stanalyst® : Une station pour l'analyse de l'information. *In:* **Actes** de Veille Stratégique Scientifique et Technologique, VSST2004, Toulouse 25-29 octobre 2004, p. 319-320.

BRACHMAN, R. J., ANAND, T. The process of knowledge discovery in databases. A Human Centered Approach. *In*: FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R. (eds) **Advances in knowledge discovery and data mining.** AAAI Press / The MIT Press, 1996. p. 37-57

BROOKES, B. The foundations of information science. Part I. Philosophical aspects. **Journal of Information Science**, v. 2, 1980. p. 125-133.

CALLON, M., COURTIAL, J-P., TURNER, W. A., BAUIN, S. From Translation to Problematic Networks: An Introduction to Co-Word Analysis, **Social Science Information,** v. 22, 1983. p. 191-235.

CALLON, M., LAW, J., RIP, A. **Mapping the Dynamics of Science and Technology**. London: The Macmillan Press, 1986.

CALLON, M., COURTIAL, J-P., PENAN, H. **La Scientométrie**. Paris : Presses Universitaires de France, 1993. (coll. Que sais-je?, 2727).

CARD, S. K., MACKINLAY, J. D., SCHNEIDERMAN, B. **Readings in information visualization using vision to think.** San Francisco, California: Morgan Kaufmann Publisher Inc, 1999.

DAILLE, B., ROYAUTE, J., POLANCO, X. Evaluation d'une plate-forme d'indexation de termes complexes, **Revue TAL**, v. 41, n. 2, 2001.

FAYYAD, U.M., PIATETSKY-SHAPIRO, G., SMYTH, P. From Data Mining To Knowledge Discovery: An Overview. *In*: FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R. (eds) Advances In Knowledge Discovery And Data Mining. AAAI Press/The MIT Press, 1996. p. 1-34.

FRAWLEY, W.J., PIATETSKY-SHAPIRO, G., MATHEUS, C. Knowledge Discovery in Databases: An Overview. *In*: PIATETSKY-SHAPIRO, G., FRAWLEY, W. J. (eds) Knowledge Discovery In Databases. AAAI Press/MIT Press, 1991. p. 1-30.

GRIVEL, L., FRANÇOIS, C. Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique, **SOLARIS,** n. 2, Presses Universitaires de Rennes, 1995. p. 81-112.

JACQUEMIN, C., DAILLE, B., ROYAUTÉ, J., POLANCO, X. In Vitro Evaluation of a Program for Machine-Aided Indexing, **Information Processing and Management**, v. 38, n. 6, 2002. p. 765-792.

POPPER, K. **Objective Knowledge**. Oxford: The Clarendon Press, 1979.

LEBART, L., MORINEAU, A., PIRON, M. **Statistique exploratoire multidimensionnelle.** Paris : DUNOD, 1997.

LELU, A. **Modèles neuronaux pour l'analyse de données documentaires et textuelles**. Thèse de doctorat de l'Université de Paris 6, 1993.

LEYDESDORFF, L. The Relations between Qualitative Theory and Scientometric Methods in Science and Technology Studies. **Scientometrics**, v. 15, n. 5-6, 1989. p. 333-347

LEVY P. **Les technologies de l'intelligence**. Paris: La Découverte, 1990.

MAIMON, O., ROKACH, L. (eds). **The Data Mining and Knowledge Discovery Handbook.** Berlin: Springer, 2005.

POLANCO, X. La notion d'analyse de l'information dans le domaine de l'information scientifique et technique [Colloque INRA, 21-23 octobre 1996, Tours]. *In* : VOLLAND-NEIL, P. **L'information scientifique et technique : Nouveaux enjeux documentaires et éditoriaux**. Paris : INRA, 1997. p. 165-172.
POLANCO, X. Plus que d'un système d'information : il s'agit de transformer l'information en connaissance et la connaissance en action, **Le Micro Bulletin**, Paris, CNRS, Délégation aux systèmes d'information, 1999. p. 15-25.
POLANCO, X., FRANÇOIS, C., ROYAUTÉ, J., BESAGNI, D. STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology. **Proceedings of the 8TH International Conference on Scientometrics and Informetrics,** July 16-20, Sydney, Australia, v. 2, 2001. p 871-873.
POLANCO, X. STANALYST. Una aplicación para nuevos estudios bibliométricos sobre bases de datos locales. II Seminario Internacional sobre Indicadores de Ciencia, Tecnología e Innovación, 16-18 de enero de 2006, Santiago, Chile.
POLANCO, X. STANALYST, un sistema de ayuda al análisis de la información. *In*: VERMEREN, P. (ed.) **El espacio público de las ciencias sociales y humanas. El papel político y los paradigmas**. Estudio comparativo Francia-América Latina. Jornadas Internacionales 2006. Buenos Aires, Editores del Puerto, 2007(a). p. 98-103.
POLANCO, X. et alii. STANALYST-SciELO: Modelo y uso para la vigilancia científica, **VII Congreso Iberoamericano de Indicadores de Ciencia y Tecnología**, RICYT-FAPESP, 23-25 de Mayo, 2007(b). São Paulo, Brasil.
RIP, A., COURTIAL, J-P. Co-word maps of biotechnology – An example of cognitive scientometrics, **Scientometrics**, v. 6, 1984. p. 381-400.
ROYAUTE, J. **Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information**. Thèse de doctorat de l'Université Henri Poincaré, Nancy 1, 1999.
SALTON G., MCGILL M. J. **Introduction to Modern Information Retrieval**. New York : McGraw-Hill Book, 1983.
SAPORTA, G. **Probabilités, analyse des données et statistique**. Paris: Editions TECHNIP, 1990.
VAN RIJSBERGEN, C. J. **Information Retrieval**. London: Butterworth, 1979.

ABSTRACT
STANALYST: Is it useful for doing what? How is it made? To which design does answer it? We will try to give answer to these three questions locating us in the context of the information science. Here we concentrate primarily on the function and the design which are at the base of STANALYST by leaving to side instructions for use. It is not a technological approach remaining exclusively at the level of an information technology description. On the contrary, the task that we assert ourselves is to place technology within the framework of a certain design of the work on information. Our central subject is to show what the information analysis means and how we can carry out this analysis with a tool designed to produce elaborated information, and still to more transform information into knowledge. The interest of the subject is even more

sensitive at the present time where our society is recognized to be an information society, and it is estimated moreover that she evolves to a knowledge society.

KEYWORDS: Technology of Intelligence. Elaborate Information. Information Analyse. Bibliometry. Informetry

Originais recebidos em: 08/02/2008
Texto aprovado em:13/03/2008