



CENTRE INTERNATIONAL DE RECHERCHE SUR LE CANCER  
PROJET POUR LE DEPARTEMENT D'INFORMATION

Document de travail

préparé par

Mme M. Wolff-Terroine,  
Chef de l'Unité de Documentation,  
Institut Gustave Roussy, Villejuif (Seine)

Le problème de l'information est un problème grave, présent dans tous les domaines de la connaissance. Partout, on assiste à un accroissement exponentiel de la littérature, partout, les chercheurs sont submergés sous des flots de publications. Les chercheurs perdent donc beaucoup de temps pour arriver à trouver, souvent incomplètement, les informations pertinentes.

Aussi, depuis quelque temps a-t-on pu assister au développement de nombreux services spécialisés chargés de la dissémination des informations auprès des personnes intéressées.

Si la question de la recherche des informations est à l'heure actuelle un problème difficile dans des domaines à limites bien définies, elle devient extrêmement complexe quand il s'agit d'un domaine multidisciplinaire tel que la recherche cancérologique. En effet, le cancer, envisagé dans son ensemble, couvre un très grand nombre de domaines extrêmement variés allant de la pathologie jusqu'aux sciences fondamentales telles que la biochimie, la virologie, l'immunologie, la radiobiologie, etc. Les chercheurs travaillant dans ce domaine ont à consulter d'innombrables revues, de nombreux index, sans être sûrs pour autant de n'avoir rien laissé échapper, car les domaines d'investigations sont variés et ne se recouvrent que partiellement les uns les autres.

Il est donc particulièrement important, dès qu'on envisage la création d'un centre international sur le cancer, d'y prévoir d'emblée un département d'informations.

Cette étude va essayer de dégager quelles peuvent être les activités d'un tel département d'information et comment on peut concevoir leur réalisation, compte des réalisations déjà existantes dans ce domaine.

#### ACTIVITE DU DEPARTEMENT D'INFORMATION

Schématiquement, le rôle d'un département d'information est de réunir les informations originales (de quelque nature qu'elles soient) et de les transformer en instruments de travail de natures diverses destinés à faciliter aux chercheurs la collecte des informations dans un domaine ou sur une question scientifique donnée.

A. Les informations collectées seraient :

- 1) Les informations publiées : articles de périodiques, livres, compte rendu de congrès, concernant le cancer sur le plan clinique et expérimental.
- 2) Les informations non publiées : le département d'information rassemble tous les renseignements sur les travaux en cours.
- 3) Les informations sur les chercheurs et les institutions travaillant les problèmes de cancérologie.
- 4) Les renseignements concernant tous les congrès, colloques, séminaires, au cours desquels les problèmes de cancérologie sont abordés.
- 5) On pourrait concevoir aussi la réunion d'informations précises permettant de répondre à des questions définies sur un sujet donné, ces réponses étant des réponses directes et non des listes bibliographiques.

Il s'agit de savoir en fonction de ces informations quels produits finis on désire obtenir; les possibilités sont nombreuses :

Index signalétique (sur titre ou sur article)

Revue de résumés

Mises au courant régulières périodiques dans les domaines d'abonnés éventuels

Réponses à des questions occasionnelles

Publications de bibliographies courantes

Publications périodiques de listes de congrès, de chercheurs, de centres de recherche.

Le choix doit être fait en fonction des besoins et des possibilités financières. (Il est à noter que, dans l'ensemble, ces choix ne sont pas exclusifs les uns des autres.) Quoi qu'il en soit, dès lors qu'il s'agit d'information dans un domaine scientifique, deux produits finis semblent essentiels : la publication d'un index signalétique rapide, la possibilité de recherche rétrospective à la demande sur un sujet donné.

#### ACTIVITES ET METHODES

La formation de ces informations en produits finis fait appel à différentes activités :

- A) La collecte des documents.
- B) L'analyse des documents pour en extraire les informations et les présenter sous la forme la plus commode en vue de l'exploitation ultérieure.
- C) Le traitement de l'information, c'est-à-dire la préparation des divers instruments de référence à partir des résultats de l'analyse précédente.
- D) La diffusion des informations, c'est-à-dire la communication des données obtenues après les opérations de traitement aux chercheurs intéressés.

Chacune de ces fonctions présente ses problèmes particuliers; dans certains cas, on se trouve devant différentes options entre lesquelles il faut choisir dès le début.

A) La collecte. Elle a deux objectifs essentiels :

- a) L'exhaustivité des dépouillements.

Certains savants sont opposés au principe de l'exhaustivité de la collecte des documents : ils estiment en effet qu'un centre de documentation spécialisé ne doit traiter que les documents les plus importants et se limiter à la collecte de renseignements provenant de certains pays seulement ou de certaines revues seulement. Cette attitude, défendable en principe, est cependant très dangereuse; en effet, il est impossible de préciser les pays et les revues produisant du travail valable, cette appréciation étant forcément entachée de subjectivité; de plus, il semble absurde de mettre

des limitations nationales à un centre international. Il nous semble donc préférable de tendre vers la plus grande exhaustivité possible de la collecte, quitte à introduire un facteur de pondération lors de l'analyse.

b) La rapidité.

Une information trop tardive a souvent perdu son intérêt pour le chercheur; la rapidité est donc un impératif primordial. On connaît les délais très longs qui séparent ordinairement la publication d'un document original et son signalement sous une forme quelconque (titre, résumé, compte rendu). Tout doit être mis en oeuvre au stade de la collecte et aux étapes ultérieures pour réduire ce délai. En particulier, on pourrait concevoir que le centre international de recherche sur le cancer soit informé par les revues ou par les auteurs de tous les travaux sur le cancer en cours d'édition. De même, le centre international de recherche sur le cancer serait informé le plus rapidement possible de tous les travaux en cours d'exécution (cf. 3ème partie).

B) L'analyse

L'analyse documentaire est l'opération qui cherche à représenter un document donné sous une forme différente de la forme originale (traduction, résumé, indexation) pour en faciliter la consultation ou le repérage par les spécialistes intéressés.

On se trouve ici en face de plusieurs options :

- a) Faut-il établir un résumé de tout document ?
- b) Faut-il indexer tout document au moyen de mots clés ou de phrases clés ? Cette indexation doit-elle être faite à partir du texte intégral, du résumé ou du titre ? Cette indexation doit-elle être faite par l'homme ou par la machine ?
- c) Faut-il traduire les titres des documents ou seulement leur formulation indexée ?

a) On ne reprendra pas ici le problème théorique de l'utilité des résumés et de leur coût. En effet, le domaine du cancer est déjà fort bien couvert par différentes revues d'abstract, principalement de langue anglaise. Ces revues, soit nationales,

telles que le Referativnii Journal, le Bulletin signalétique du CNRS, soit internationales, telles que Excerpta Medica, soit axées sur un sujet bien délimité, telles que Carcinogenesis Abstracts, Cancer Chemotherapy Abstracts, Leukemia Abstracts, Berichte über die Gesamte Biologie, Biological Abstracts, etc., sont dans l'ensemble toutes de haute qualité. Il y aurait donc là un double emploi manifeste et il paraît nécessaire d'écarter a priori l'option d'une revue de résumés.

b) Indexation de tous les documents. Quels que soient les progrès techniques, il n'est pas encore question d'enregistrer en mémoire le texte intégral d'un document et, même à ce moment-là, il faudra encore caractériser ce texte avec des concepts normalisés. Donc l'indexation, quelque coûteuse qu'elle soit, est une étape indispensable pour la mise en mémoire d'un document.

1) Indexation mécanique. Cette technique a pour objet de repérer automatiquement dans le titre des documents certains termes considérés comme importants : on constitue ensuite des sortes de tables où ces documents sont rangés par mots vedettes, par auteurs, par collections ... (cf. KWIC, index). Cette indexation est assez facilement réalisable car le "programme" correspondant peut être acheté. On peut, à juste titre, faire plusieurs objections à cette méthode; en particulier, elle ne repose que sur la lecture du titre et par conséquent reste souvent à la surface des informations (cf. essai statistique fait à l'IGR sur 1000 documents concernant le cancer du sein et du poumon); de plus, la consultation de tels index est assez rébarbative.

2) L'indexation manuelle est la représentation du contenu d'un texte par une suite de mots clés ou de phrases clés empruntés à un lexique ou à un langage artificiel spécialement conçu pour le traitement de l'information dans un domaine scientifique donné. La création d'un tel lexique adapté à la recherche cancérologique est un travail difficile. En effet, les différents lexiques existants sont beaucoup trop généraux pour pouvoir être utilisables (cf. communication du M.D. Anderson Hospital au 26th Annual Meeting of American Documentation Inst., Chicago, 1963). Un seul lexique à notre connaissance existe, c'est celui de l'Institut Gustave Roussy qui, complété par un thésaurus, est opérationnel depuis deux ans.

On pourrait le compléter par le thésaurus encore en voie d'élaboration au M.D. Anderson Hospital en radiobiologie. Les méthodes d'indexation demandent donc à être précisées tant au point de vue des problèmes lexicaux que des problèmes syntactiques (cf. infra plan de réalisation).

c) La traduction peut être envisagée à divers moments : traduction du titre ou traduction des mots clés. Les solutions au problème de la traduction demandent à être envisagées d'un point de vue pragmatique. En théorie, il est parfaitement possible de traduire tous les titres et tous les termes indexés. En pratique, ceci est difficilement applicable. En effet, indépendamment même des facteurs financiers (les traducteurs spécialisés sont rares et chers), si l'on admet que la rapidité est un élément essentiel du bon fonctionnement d'un tel département d'information, il faut renoncer à la traduction en plusieurs langues et adopter une seule langue de travail. Cette langue de travail semble, dans l'état actuel des choses, devoir être l'anglais. Il faudrait donc prévoir la traduction systématique de tous les titres et l'indexation directe des articles en langue anglaise. (Par contre, les publications du département d'information n'ayant pas trait à des informations scientifiques sensu stricto (répertoire de chercheurs, listes d'établissements scientifiques, programmes de congrès...) pourraient être publiés dans les langues officielles du centre international.

### C) Traitement

Toutes les informations après analyse doivent être codées pour pouvoir être mises en mémoire, et ensuite explorées lors de leur exploitation - (cette exploration consistera le plus souvent à comparer les termes d'une "question" et les termes trouvés dans la représentation indexée des documents en mémoire). Dans l'état actuel des techniques d'enregistrement de l'information, il semble nécessaire de prévoir d'emblée une automatisation du stockage et de l'exploration des informations (cf. II Moyens).

A noter que, dans le cas des références bibliographiques, il faudrait prévoir la mise en mémoire non seulement de la représentation indexée et codée de chaque document, mais aussi de sa référence bibliographique et de son titre complet : on éviterait ainsi d'obtenir à la sortie de la tabulatrice une liste de numéros de documents; cette

méthode fréquemment employée oblige à des consultations pénibles pour obtenir dans une deuxième phase l'identification des numéros de documents pertinents extraits et leur référence bibliographique détaillée.

#### D) Diffusion

Les divers modes de diffusion ont déjà été énumérés plus haut, index signalétique, service "SVP", questions occasionnelles ou abonnements à des bibliographies particulières ... Mais il faut prévoir deux autres types de services indépendants des précédents :

- la reproduction matérielle des documents originaux,
- la traduction éventuelle des documents.

### MOYENS

#### Equipements

Les problèmes d'équipement ne se posent qu'à partir de la phase d'indexation et de codification.

Les appareillages évoluant extrêmement rapidement, il serait prudent de prévoir une mise en mémoire permettant le passage simple d'un équipement initial d'étude à un équipement plus lourd.

Dans l'état actuel des choses, l'indexation réellement automatique étant encore en phase d'étude (les index types KWIC étant plutôt des tris et des tabulations que des "index" réels), une indexation manuelle semble devoir être envisagée, l'automatisation ne commençant qu'après la frappe dactylographique d'une fiche manuelle d'indexation. On prévoirait une entrée en mémoire par cartes ou plutôt par bande perforée. Il serait installé un atelier d'exploitation automatique comportant notamment un matériel mécanographique "classique", une machine à bande perforée avec lecteur auxiliaire et perforatrice auxiliaire, et on louerait les services d'un calculateur de moyenne puissance dans une première phase d'étude.

Dans l'ensemble, pour les travaux d'imprimerie, l'importance du centre ne semble pas justifier l'utilisation de machines permettant la composition directe à partir des états fournis par les tabulatrices. On fera plutôt effectuer par un sous-traitant des reproductions en offset des états des renseignements fournis par l'imprimante.

Il y aura lieu, cependant, de prévoir un offset de bureau pour les publications à faible tirage.

Des appareils de reproduction sont nécessaires et doivent être choisis dans une gamme très étendue (photocopie, xerographie ...).

#### Techniciens

Le personnel à prévoir pour remplir ces différentes tâches est assez divers.

Pour la collecte, il est nécessaire de prévoir un bibliothécaire médical et scientifique.

La lecture des informations ne peut être faite que par du personnel médico-scientifique compétent. Ce personnel, pour être réellement compétent, doit être au courant des problèmes. Il est difficile de concevoir des médecins travaillant à plein temps pour un département d'information, car ils deviendraient très rapidement trop éloignés des problèmes du moment.

De plus, il faut prévoir de nombreuses spécialités : en effet, il faut des biochimistes pour apprécier un article en biochimie et des virologistes pour apprécier un article en virologie et non point du personnel ayant une bonne culture générale médicale. On demandera à ce personnel spécialiste, engagé lui-même dans la recherche, de faire une analyse descriptive des documents, mais aussi une appréciation des documents, basée sur les critères suivants : apport de faits nouveaux, apport de théories nouvelles, revue d'ensemble, article de vulgarisation.

Pour le traitement des informations, il faut au contraire prévoir du personnel ayant une solide culture médicale et scientifique de base et entraînés aux méthodes de traitement de l'information et du personnel technique de mécanographie.

Quant à l'exploration des informations mises en mémoire, elle requiert la collaboration des analystes spécialisés et des techniciens de l'information.

Si, par contre, l'on envisage d'emblée l'acquisition d'un ordinateur pour le centre, il faut prévoir que tous les ordinateurs ne se prêtent pas au même degré aux opérations de documentation automatique. Sans vouloir entrer ici dans des considérations techniques, il faut préciser que le choix devrait s'orienter parmi les machines électroniques alors disponibles vers celles qui présenteraient alors les qualités suivantes :

- mémoire se prêtant à une organisation par groupes d'unités individuellement adressables,
- temps d'arrêt très court,
- vitesse de transfert élevée à l'intérieur de la machine,
- imprimante riche en types de caractère.

Dans ces conditions, il faudrait évidemment prévoir des spécialistes capables d'assurer la bonne marche d'un centre de calcul, l'élaboration et le testage des programmes.

#### ORGANISATION ET STRUCTURE

La collecte des informations. Le département d'information possède une vaste bibliothèque recevant les principaux périodiques.

D'autre part, il établit avec les instituts de recherches un réseau de relations. Ces instituts de recherches deviennent ses correspondants et lui envoient les documents qu'il ne recevrait pas sur place.

Grâce à sa liaison avec les divers instituts, il est tenu au courant des travaux en cours, du nom et de la qualité des chercheurs, etc.

L'indexation est le plus possible centralisée. A première estimation, le centre devrait trouver dans son entourage scientifique immédiat des spécialistes capables

d'analyser environ les deux tiers des informations. Le dernier tiers qu'il serait impossible d'analyser au centre pour raison de spécialité ou pour raison linguistique serait confié à des correspondants; ces derniers enverraient au centre une sorte de résumé télégraphique qui serait indexé au centre. (On pourrait concevoir qu'après une longue phase de rodage, les correspondants puissent indexer eux-mêmes : Viniti a attendu 5 ans avant de tenter un essai analogue !)

L'indexation donc, le traitement et l'exploration automatique des documents doivent être centralisés.

En effet, l'indexation est une tâche fort particulière qui requiert l'apprentissage d'un langage documentaire pourvu de règles propres.

La codification, le traitement et l'exploration étant par excellence les phases mécaniques de l'exploitation doivent être exécutés là où se trouve l'équipement. Notons cependant que l'exploration ne peut se faire de façon satisfaisante, en particulier dans le cas de réponses à des questions occasionnelles, qu'en étroite liaison avec les spécialistes du domaine considéré, et ceci, quels que soient les raffinements lexicaux et syntactiques prévus a priori. (En effet, quel que soit le degré de normalisation de l'indexation, la correspondance est rarement univoque entre l'énoncé d'une question et la représentation des documents idéalement visés. L'utilisation d'un thésaurus permet par ses extensions verticales et horizontales la comparaison entre "questions" et "documents". Cependant, en pratique, fréquemment, seul l'avis du spécialiste dans le domaine considéré permet de fixer une limite à ces extensions, ces limites étant essentiellement mouvantes dans un domaine de recherche "en pointe".) Une fois de plus s'affirme la nécessité pour un centre d'information spécialisé de l'étroite collaboration avec les divers spécialistes.

#### La diffusion

La publication d'un index signalétique, c'est-à-dire d'une liste de références classées, ayant pour objectif essentiel la rapidité, pourrait être une publication hebdomadaire ou bimensuelle. On a admis que le centre est capable d'indexer lui-même les deux tiers des documents et qu'un tiers seulement est analysé par des correspondants;

la plus grande partie des documents étant indexée au centre en vue de leur signalisation et de leur mise en mémoire, les délais sont ainsi relativement faibles; ceci associé à une publication de périodicité fréquente (hebdomadaire ou bimensuelle) permettrait d'obtenir une indication rapide des nouvelles publications.

Si cette solution n'est pas retenue, il faudrait alors prévoir un double système d'indexation : une première indexation en surface par les employés du centre pour les besoins de la signalisation rapide, et une deuxième indexation en profondeur par les spécialistes des domaines considérés; cette deuxième solution permettant peut-être une rapidité légèrement supérieure, mais étant moins spécifique, plus lourde et plus coûteuse.

Quant aux demandes de renseignements, elles demandent deux types différents de diffusion :

- les réponses à des questions occasionnelles précises,
- des abonnements particuliers pour mises au courant périodiques sur un sujet précis.

D'autres publications doivent être envisagées (listes d'institutions, de congrès, etc., cf. supra); leur périodicité demande à être précisée.

#### PLAN DE REALISATION

La création d'un tel département d'information est une opération demandant de sérieuses études préliminaires, méthodologiques et d'organisation. Pour éviter un gaspillage de temps, d'effort et d'argent, il y aurait sans doute lieu d'utiliser largement dans des limites qui restent à préciser les réalisations existant dans ce domaine. A l'heure actuelle, un seul centre pratique de façon opérationnelle, une mise en mémoire systématique et une exploitation mécanique des informations (publication d'une revue signalétique, "Information retrieval", à la demande), c'est l'Institut Gustave Roussy à Villejuif, France, qui a construit un lexique et établi un thésaurus; il reste cependant insuffisant pour la partie biochimie, actuellement en voie d'élaboration. Le M.D. Anderson Hospital, à Houston, étudie, sur un nombre de documents limités, la création d'un thésaurus dans le domaine de la radiobiologie en rapport avec le cancer. Le Chester Beatty, à Londres, publie une liste mensuelle classée indicative des nouvelles parutions.

Le National Cancer Institute prépare une classification extrêmement détaillée de tous les travaux effectués grâce à des fonds fournis par lui. Il existe évidemment de nombreuses autres institutions pratiquant une indexation et une recherche de documents concernant le cancer, mais leurs travaux ne sont pas axés sur le cancer et le vocabulaire utilisé manque toujours de spécificité.

Comment donc pourrait être prévues les études préliminaires et les débuts de réalisation compte tenu des travaux déjà existants.

### Collecte

Les responsables du futur département d'informations commenceraient à établir :

- 1) la liste des périodiques nécessaires à la bibliothèque du centre,
- 2) un répertoire des bibliothèques des grands centres engagés dans la recherche cancérologique,
- 3) un répertoire permanent des centres, laboratoires, chercheurs. On pourrait mettre à profit les recensements établis par l'OMS dans ce domaine,
- 4) un répertoire permanent des recherches en cours (voir le travail effectué par l'OMS).

Ces instruments d'information générale devront être constitués et exploités par des procédés mécanographiques.

### Analyse

Création d'un lexique et d'un thésaurus. Ce travail fondamental de très longue haleine, pourrait en grande partie être évité par l'utilisation du thésaurus de l'Institut Gustave Roussy. Certains détails gagneraient à être précisés. Il faudrait accélérer la constitution d'un thésaurus en biochimie; on pourrait utiliser l'expérience de M. D. Anderson en radiobiologie.

Il serait intéressant de prévoir une classification simple pour la présentation de l'index signalétique; on pourrait utiliser celle de l'Institut Gustave Roussy complétée par celle du Chester Beatty.

Des études méthodologiques doivent aussi être entreprises pour tester la valeur des méthodes d'indexation : utilisation d'une simple indexation coordonnée ou introduction de facteurs syntactiques simples ou complexes - on pourrait peut-être utiliser les études actuellement en cours dans ce sens à l'Institut Gustave Roussy. En effet, des études statistiques y sont actuellement poursuivies pour étudier la valeur du thésaurus et la validité des différentes méthodes d'indexation dans les domaines considérés.

De plus, il serait nécessaire de prévoir un recensement des spécialités couvertes afin de pouvoir établir les domaines couverts par le département d'information du centre international et ceux pour lesquels il faudra prévoir des correspondants.

#### Traitement

Au minimum, il faudra prévoir pour la phase initiale une machine type Flexowriter avec perforateur auxiliaire et un atelier mécanographique classique. Dans cette phase, un ordinateur à demeure est inutile - on pourrait, si nécessaire, utiliser les services de travaux à façon faits par les grands fabricants d'ordinateurs. Pour des raisons d'économie et de pratique, il pourrait peut-être être utile, dans une phase d'étude, d'utiliser le matériel de centres déjà existants et habitués à traiter des informations.

A ce stade du travail, il faudrait étudier le matériel nécessaire pour les étapes ultérieures. Si l'acquisition d'un ordinateur ayant les capacités spécifiées plus haut n'est peut-être pas concevable dans le cadre de l'institut international, il faudrait voir le problème d'un traitement éventuel des données par une institution ayant le type d'ordinateur désiré (OMS, IGR, Chester Beatty) et mettre au point une programmation appropriée.

- Diffusion

Là encore, il faudrait étudier les formes des publications faites par le département d'information, choisir entre la création "de novo" d'index signalétique ou l'amplification de services déjà existants.

Le matériel de reproduction et d'impression devrait être étudié et acheté.

Cette étude préliminaire demanderait suivant les points envisagés (et les crédits disponibles) un à deux ans.

Il est presque impossible d'établir un budget chiffré même très approximatif pour un tel département d'information. En effet, comme il a été dit de nombreuses options, tant théoriques que pratiques, sont en présence et rendent les estimations difficiles. Aussi la note annexe<sup>1</sup> de ce rapport est-elle faite sous toutes réserves.

La création d'un tel département d'information est donc un problème délicat quant à la conception et à la réalisation. Quelles que soient les décisions prises et les options choisies, deux points semblent primordiaux pour la qualité du futur travail effectué :

- une étude approfondie des problèmes lexicologiques,
- le recrutement de spécialistes réellement compétents, et eux-mêmes engagés dans la recherche, pour effectuer le travail d'analyse.

Les impératifs méthodologiques et intellectuels étant ainsi remplis, un tel département, muni d'un bon équipement matériel, serait de la plus haute utilité à tous ceux qui sont engagés dans la lutte contre le cancer.

---

<sup>1</sup> Elle sera remise séparément.