Deuxiême colloque du chapître de l'International Society for Knowledge Organization, ISKO99, Lyon, France, Oct 21-22, 1999.
Draft English version http://people.ischool.berkeley.edu/~buckland/lyon.html

Forme, Signification, et Structure des Systèmes de Sélection du Savoir

Michael Buckland, Professor,
School of Information Management and Systems,
University of California, Berkeley, CA, USA

"*Une bibliographe contemporaine soucieuse de clarté a lancé cette brève définition: Un document est une preuve à l'appui d'un fait.*" (Suzanne Briet, 1951, 7)

Mesdames, Messieurs, l'invitation à ce colloque d'ISKO France est un très grand honneur. Je vous remercie et je dois reconnaître l'influence dans ma vie professionelle des oeuvres de documentalistes francophones, surtout Suzanne Briet et Paul Otlet. On m'a invité à faire une allocution sur les origines, le développement historique, l'état actuel, et l'avenir de l'indexation. À la place d'un tel discours qui, évidement, serait beaucoup trop ambitieux pour moi, j'ai voulu présenter quelques idées touchant les systêmes de sélection. Quels phénomènes intéressent les chercheurs en systèmes d'organisation du savoir? Qu'est-ce qu'on peut dire sur la structure de tels systêmes? Comment peuvent les usagers de tels systêmes comprendre ce que les données, les documents, et les metadonnées signifient? Comment pouvons nous joindre en une seule perspective la technologie et la signification? En bref, comment caractérisons nous les systêmes d'organisation du savoir. Les remarques que je vais présenter aujourd'hui proviennent d'idées développées avec mes collègues à Berkeley.

Mon titre est "Forme, Signification, et Structure des Systèmes de Sélection du Savoir". Le mot "Forme" concerne les phénomènes d'intérêt ("Information-as-thing").
Le mot "Signification" est susceptible de désigner tantôt la faire (la signification comme procès: "Information-as-process"), tantôt l'état (ce qui est signifié "Information-as-knowledge"). J'utilise "Structure" pour représenter le génie, le systême et le savoir-faire du documentaliste.

1. De quelles choses nous occupons nous?

Les membres d'ISKO s'intéressent à la structure du savoir. Cependant, quand nous utilisons n'importe quelle espèce de technologie pour développer des systèmes opérationels, nous ne nous occupons plus directement de conceptions abstraites, mais de données, de textes, et d'autres objects concrets. La technologie est nécessairement matérielle. Donc nous nous occupons indirectement de savoir. Nous nous occupons directement de signes, de représentations de la connaissance, d'objets que nous considérons comme significatifs. On pourrait dire que nous nous occupons de documents, mais de document dans n'importe quelle forme. Les documents ne sont pas seulement faits de texte.

Parler de "document" de cette façon n'est pas original. En 1937 l'Institut International de Coopération Intellectuelle, une organisation crée par la Société des Nations, a collaboré avec l'Union Français des Organismes de Documentation, à fin de définir des termes techniques,

"document" y compris:

Document : Toute base de connaissance, fixée matériellement, susceptible d'être utilisée pour consultation, étude ou preuve. Exemples: manuscrits, imprimés, représentations graphiques ou figurées, objets de collections, etc...

Document : Any source of information, in material form, capable of being used for reference or study or as an authority. Examples: manuscripts, printed matter, illustrations, diagrams, museum specimens, etc.... (Anon. 1937: 234)

2. L'Indexicalité

Suzanne Briet (1894-1989), bibliothècaire, documentaliste, historienne, a avancé le concept de "document" en 1951 dans son manifeste intéressant *Qu'est-ce que la documentation?* Elle déclare, tout d'abord, que "Un document est une preuve à l'appui d'un fait" (Briet, 1951, 7). Ensuite, elle explique qu'un document est:

"*...tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène ou physique ou intellectuel*." (Briet, 1951, 7).

Par consequence on ne peut pas considérer que le métier de documentaliste (ou bien "Information Management") s'occupe de textes, mais, plutôt, de toute espèce de preuve, de témoignage, d'évidence et que cette preuve ("le document") est de forme concrète et non pas abstraite. Remarquons que Briet a employé le mot "indice." À mon avis, le mot "indice" veut dire qu'un objet ne devient une preuve (un document) que si on a placé cet objet en rapport avec des autres preuves (des autres documents). C'est à dire que les documents doivent être arrangés "indexicalement," les uns avec les autres.

Une approche plus contemporaine serait de dire que le sens est construit par le spectateur. Que tout objet pourrait, dans certaines situations, être preuve, être un document. Donc tout objet *peut devenir* signifiant. Tout object concrèt peut être un document. C'est possible, même si c'est peu vraisemblable. Malgré tout, nous retenons deux suppositions de Briet: Que tout objet peut être un document; et que l'essence de la documentation est d'arranger volontairement ces objets dans des relations indexicales. Ces relations sont, bien sur, d'un intérêt tout particulier pour les membres de l'ISKO.

3. Que faire avec des documents?

En plus d'être créés, les documents sont sélectionnés, représentés, et utilisés.

3.1. Sélection.

Nous rassemblons des documents dans des collections et nous extrayons des documents de nos collections. D'habitude on crois que ce sont des procédés différents. Mais le développement de collections et l'extraction de documents d'une collection sont tous deux des procédés de seléction. Dans l'un et dans l'autre un ou plusieurs documents sont accordés un rang privilégés vis à vis d'autres documents. En Anglais, on parle de "information retrieval systems" et de "search engines." Bien sûr on cherche et on extrait, mais il y a aussi un élément de choix. Moi,

j'aime la terminologie des années 1930: "Machines à sélectionner".

3.2. Représentation.
Nous créons des abrégés, des fiches, et autres représentations descriptives de documents. Ces représentations peuvent servir comme substitut du document original. Nous créons des metadonnées (metadata) qui décrirent les données (data). En fait, nous faisons des représentations bibliographiques. Il y a une continuité entre la très courte entrée d'un index, et la version complète d'un document comprenant une description bibliographique extensive. À la base, nous dérivons de nouvelles représentations à partir de documents existents.

3.3. Usage
Il est difficile de prédire l'usage de documents. D'habitude nous ne savons pas qui va utiliser un tel document. Peut-être que personne ne l'utilisera. Normalement nous ne savons pas si un document a été lu soigneusement ou examiné superficiellement -- ou bien si on a examiné le document entier ou consulté une petite partie. Surtout nous ignorons les conséquences intellectuelles ou practiques de cet examen du document. La consultation des metadonnèes se dèroule comme la consulation des données (d'un document), sauf que les metadonnées sont bien brèves et qu'il y a moins de base pour comprendre l'intention de l'auteur (l'indexeur).

4. Structure.

Regardons des exemples de sélection et de représentation.

4.1. L'indexation automatique. Avec les documents numériques on peut utiliser des systêmes diverses d'indexation automatique par logiciels. Le systême KWIC arrange chaque mot du document (avec ses mots adjoints pour offrir un contexte) en liste alphabetique. Chaque ligne d'un index KWIC décrie un toute petite partie du document. La totalité des lignes KWIC dérivées du document constitue un représentation mechanique du document entier. Les systêmes sous forme vectorielle (par exemple SMART) créent des représentations mechaniques plus complexes.

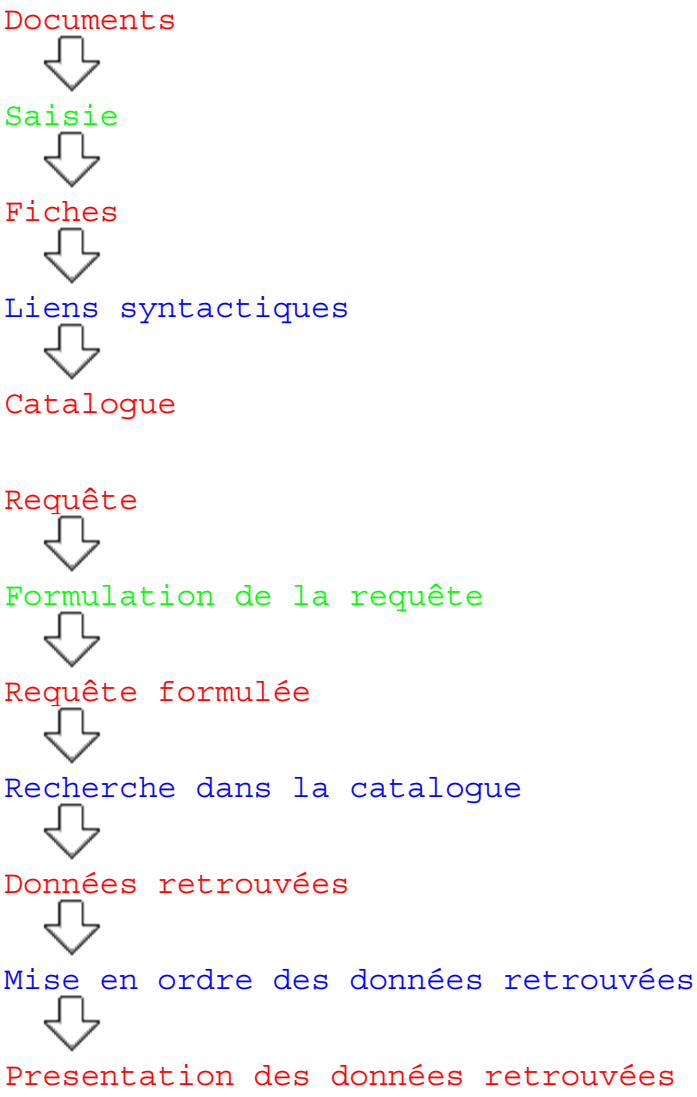
4.2. Un catalogue de bibliothèque. Considérons la creation des fiches bibliographiques soit pour un fichier ou bien comme données pour un catalogue online. On assemble les données bibliographiques qui représentent un livre. Quelques données sont dérivées du document même: le titre; le nom de l'auteur; l'année d'édition). D'autres données sont dérivées d'autres sources (par exemple un bibliographie nationale, un thesaurus). Une fiche de catalogue est une représention d'un livre, suivant les techniques et régles conventionelles à l'égard de la forme, du contenu, et de l'origine du livre (e.g. ISBD, AACRII, LCSH, LCC).

Cependant, les metadonnées crées par le documentaliste peuvent être modifiées, voire normalisées, par les liens syntactiques imposée par l'éditeur du catalogue: EM (Employer); EP (Employé pour); etc, et l'harmonisation des noms personels.

En parallèle, l'usager de la bibliothèque veut poser un question. Il est probable que sa requête ne correspondra pas exactement à la forme des entrées du catalogue, et donc qu'une version modifiée de cette requête devra être dérivée afin d'y correspondre plus précisément.

Il faut donc que la terminologie de la question s'accorde avec la terminologie du cataloge. Dans le mesure où une correspondance existe, l'ensemble de documents pertinents est obtenue. De plus, les cataloges automatisés arrangent généralement les resultats obtenus en ordre alphabétique avant de présenter les données.

La figure suivante montre ce procédé.



Nous pouvons observer que le procédé contient une suite d'opérations (vertes et bleues) et que chaque opération produit un nouvel ensemble (rouge) d'objets à partir d'un ensemble (rouge) d'objets précédents. Notons que nous pouvons regarder le catalogue entier comme un document complexe qui représent la collection entière. En même temps, si nous considérons un niveau moins agrége, nous pouvons regarder le catalogue comme un ensemble de fiches individuelles (elles-mêmes des petits documents) chacune représentant un livre dans l'ensemble de livres qui constituent la collection. Chaque index (auteur; titre; mot-clef; classification) est un sous-ensemble du cataloge. *On trouve que chaque étape produit un nouvel ensemble d'objets à partir d'un ensemble précédent à travers quelques opérations et qu'il n'y a que deux catégories*

d'opérations.

1. Une catégorie d'opérations (bleues) arrangent (misent en ordre, partitionent) les membres d'un ensemble. Nous comprenons dans cette catégorie un ordre total (strictly ordered set), des ensembles faiblement ordonnés (weakly ordered set) - surtout l'ordre binaire à deux sous-ensembles: des données retrouvées et des données non-retrouvées -- et aussi la combinaison de deux ensembles pour fair un sur-ensemble.

2. L'autre catégorie d'opérations (vertes) comprend les transformations qui modifient des membres d'un ensemble. La dérivation des fiches (ou bien des lignes KWIC ou représentations vectorielles) des documents originaux est de cette seconde catégorie.

Ces deux catégories d'opérations sont aussi les deux espèces d'activités que nous avons notés auparavant dans la section "Que faire avec les documents?": La sélection; et la création de représentations (ou versions) de documents. Si on analyse les systèmes bibliographiques, de recherche d'information, et de filtrage, on trouve toujours cette structure: *une châine d'opérations sur des ensembles de données qui produisent toujours un ensemble nouveau, soit (re)ordonné, soit transformé, sans exception*. Pour le moins, dans les enquêtes du Dr Christian Plaunt et moi-même, nous n'avons trouvé aucune exception jusqu'ici. It parâit que ce formule caracterise tous systême opérationel pour l'organisation du savoir. (Buckland & Plaunt 1994; Plaunt 1997).

5. Plusieurs Vocabulaires Co-existent

Tout système à sélectionner du savoir inclue de multiples vocabulaires. Même dans des cas primaires, par exemple quand un texte non-édité est parcouru avec une requête non-éditée, il y a au moins deux vocabulaires:

1. Le vocabulaire de l'auteur du document - ou bien les vocabulaires de plusieurs auteurs; et

2. Le vocabulaire du chercheur.

Dans les systêmes opérationnels actuels, on trouve, d'habitude, beaucoup de vocabulaires simultanés. Dans un catalogue de bibliothèque, par exemple, on trouverait trois autres vocabulaires:

3. Le vocabulaire d'indexation du documentaliste, qui modifie ou supplemente le vocabulaire de l'auteur.

4. Les liens syntactiques -- EM (Employer); EP (Employé pour); etc. -- pour harmoniser ou corriger les vocabulaires des documentalistes;

5. Le vocabulaire du chercheur tel que formulé dans une requête.

En bref, *il y a toujours des vocabulaires multiples en jeu*. L'espoire que tous ces vocabulaires soient indentiques ou se harmoniseraient est malheureusement futile.

Si on regarde plusieurs systèmes de sélection, les vocabulaires différents foisonnent! Voici un exemple: J'ai voulu chercher des livres et des articles sur "Coastal pollution" (La pollution des côtes marines) dans MELVYL, le catalogue online de l'University of California, et MEDLINE. Ni l'un, ni l'autre utilisent la phrase "Coastal pollution" et une recherche booléenne avec "coastal" et "pollution" n'a rien trouvé, malgré que des documents pertinents existaient dans les deux systèmes.

Dans le cataloge MELVYL utilisant le *Library of Congress Subject Headings*: On a du chercher sous: Marine pollution; et ensuite: Coastal zone management; Water -- Pollution; Petroleum industry and trade; Beach erosion; Coasts; Barrier islands; Coastal changes; etc.

Mais dans MEDLINE, utilisant *MeSH*, on a du employé Seawater, et ensuite: Water pollution; Bacteria; Water microbiology; Air pollution; Environmental monitoring; Bathing beaches; Environmental pollution; etc.

Remarquez la variété et le peu que les deux listes ont en commun. Ces termes d'indexation sont, certes, justifiables, mais qui pourrait possiblement en imaginer la moitié? Ici nous avons rencontré trois vocabulaires différents: LCSH, MeSH, et la mienne.

6. Correspondances et associations entre vocabulaires

C'est precisement à cause de cette multiplicité de vocabulaires, qu'il y a toujours la possibilité d'une incompatibilité lors de la transition entre vocabulaires, d'une dissonance de sens. Un chercheur peut employer le terme A et un auteur a employé le terme B. Ils peuvent vouloir indiquer le même sens -- des synonymes. Cependant, il est possible que tous les deux aient employé le terme A pour indiquer des sens différents -- des homographs.

Les vocabulaires intermédiaires (que ce soit celui du documentaliste, une requête formulée, ou la structure syndétique) peuvent être considéréés comme visant à normaliser l'usage des termes afin de rectifier toutes les discordances. L'index du documentaliste rectifie le titre donné par l'auteur en représentant le sujet du document à travers un vocabulaire standardisé. Les chercheurs expérimentés savent comment modifier leurs requêtes ou celles des autres d'une façon telle que le système y répondra utilement.

Il y a autant de re-représentations que de transitions d'un vocabulaire à un autre. Chacune de ces re-représentations présente une opportunité pour rectifier les dissonances entre le chercheur et le document, mais offre aussi la possibilité à de nouvelles dissonances d'émerger. Un bon intermédiaire de recherche (humain ou informatisé) pourrait en savoir assez pour demander un changement de terminologie et l'adaptation du vocabulaire du système.

7. Une Definition de "Vocabulaire"

Nous avons parlé de "vocabulaire" comme si c'était un langage ordinaire. Mais si nous croyons que le concept de vocabulaire est important pour les systême à sélectionner du savoir, il nous faut une définition technique, précise, et suffisante dans le métier de documentaliste.

L'*Oxford English Dictionary* (1989, vol 19, 721) offre quatre définitions de "Vocabulary":

1. Une ensemble ou liste de mots avec des explications brèves de leur sens;

2. L'étendue du langage d'une personne, classe, métier ou autre.

3. La totalité ou agrégation des mots composant une langue; et

4. Figurativement, un ensemble de formes artistiques ou stylistiques, techniques, mouvements, etc., à la disposition d'une personne particulière, etc.

La notion fondamentale est que "vocabulaire" dénote un énumeration de différentes formes d'exprimer du sens, le répertoire des formes représentatives. C'est à dire que le répertoire de termes d'indexation est le vocabulaire du documentaliste.

Les metadonnées comme langue

Dans les système d'indexation les termes sont maintes fois des adaptions plus ou moins artificielles de la langue quotidienne (par exemple: *God -- Knowableness -- History of Doctrines -- Early Church, ca. 30-600*) ou emploient une notation artficielle (par exemple "330" signifie "Sciences Economiques" dans la Classification Décimale de Dewey. Ce sont des systêmes pour coder le savoir. Evidemment chacun est une espèce de langage. Décrire est une activité de la langue. On a reconnu depuis longtemps que les systêmes d'indexation sont des langues. On parle aujourd'hui de "metadonnées," mais avant "metadonnées" on parlait de "langues documentaires," "langues d'indexation," ou bien "metalangues. (Citons Maurice Coyaud, 1966).

On peut, donc, utiliser le mot "vocabulaire" pour dénoter le repertoire de n'importe quelle langue documentaire: Les termes d'un thesaurus; les nombres d'une classification; les vedettes-matières; les valeurs d'une catégorisation. Dans le cadre de la documentation on peut employer le terme "vocabulaire" pour dénoter le répertoire de n'importe quelle champ MARC ou toute autre forme d'ensemble de metadonnées. C'est un concept puissant parce que, comme je le disaist, toute espèce de systême à sélectionner du savoir comprend une châine d'opérations sur des ensembles de données qui produisent toujours un ensemble nouveau. Le répertoire de chaque ensemble successif est le vocabulaire de cette ensemble. Ainsi ce concept de vocabulaire est devenu un concept tout à fait central.

9. Les langages humains tendent à être imprécis

Les vocabulaires tendent à être imprécis pour deux principales raisons:

1. Il y a, peut-être un manque de familiarité. Les termes sont peu connus. Qui sait que pour trouver des documents touchant les autmobiles, it faut chercher sous "TL 205" dans la classification de la Library of Congress Classification, sous "180/280" dans la classification des brevets de l'U.S. Patent Office, et sous "3711" dans la Standard Industrial Classification? Evidemment un index ou dictionnaire reliant notre langue quotidienne à chaque langue documentaire serait très utile. (Buckland with others 1999 [Dlib]). Mais un index est aussi nécessaire quand un langage "naturel" est utilisé pour la classification. Dans les statistiques officielles de commerce international des Ètats-Unis, on ne trouve pas de commerce en "automobiles," qui n'existent pas dans leur indexation en langue naturelle. On trouve des données statistiques si on cherche sous "cars", mais ces donn´es sont pour les wagons de cheimin de fer!

Il faut chercher les automobiles chez "Passenger motor vehicles, spark ignition engine," un terme d'indexation assez descriptif, mais inattendu.

2. Par ailleurs, la terminologie reste imprécise parce que l'emploi des mots est dynamique. La langue est une chose vivante. Ce qu'un mot signifie change.

À mon avis

Si l'analyse que j'ai présenté est correcte les conséquences sont nombreuse et importantes. Les systêmes d'organisation du savoir sont à la base des systêmes de langues, de vocabulaires, et ce qui signifierait qu'ils seront toujours imprécis.

Les théories formelles de l'information, qui utilisent la logique, l'entropie, le calcul de l'incertitude, et qui sont appréciées et prestigieuses au sein des sciences de l'information, resteront toujours incompletes, utiles, peut-être, mais inachevées.

La plupart des recherches de "digital libraries" s'occupent de questions d'infrastructure, et non pas de problêmes centraux pour les systêmes d'organisation du savoir.

Ces problèmes centraux concernent la langue, la représentation, l'explication semiotique de "...tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène ou physique ou intellectuel."

References

Anon. (1937). La terminologie de la documentation. *Coopération Intellectuelle* 77, 228-240.

Buckland, M. K. 1999. The Landscape of Information Science: The American Society for Information Science at 62. *Journal of the American Society of Information Science* 50, no 11 (1999):970-974. http://www.sims.berkeley.edu/~buckland/asis62.html

Buckland, M. K. 1999. Vocabulary as a Central Concept in Library and Information Science. In *Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3, Dubrovnik, Croatia, 23-26 May 1999*. Ed. by T. Arpanac et al. Zagreb: Lokve, pp 3-12. http://www.sims.berkeley.edu/~buckland/colisvoc.htm

Buckland, M. K. 1997. What is a "document"? *Journal of the American Society for Information Science* 48, no. 9: 804-809. http://www.sims.berkeley.edu/~buckland/whatdoc.html Similar text in *Document Numérique* (Paris) 2, no. 2 (1998): 221-230.

Buckland, M. and others. 1999. Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies *D-Lib Magazine* 5 (1) January 1999. Online at: http://www.dlib.org/dlib/january99/buckland/01buckland.html

Buckland, M. K. & C. Plaunt. 1994. On the Construction of Selection Systems. *Library Hi Tech* 12:4:15--28. http://www.sims.berkeley.edu/~buckland/papers/analysis/analysis.html

Coyaud, M. 1966. *Introduction a l'étude des langages documentaires*. Paris: Klincksieck.

Plaunt, C. 1997. *A Functional Model of Information Retrieval Systems and Processes*. Ph.D. dissertation, School of Information Management & Systems, University of California, Berkeley.