Indexação documentária: uma forma de representação do conhecimento registrado[1]

Virgínia Bentes Pinto[2]

A indexação documentária, atividade que se ocupa em estabelecer a representação do conhecimento registrado, faz parte de um sistema global: o sistema de recuperação de informação -SRI- o qual é constituído por um conjunto de atividades que contemplam desde o processo de seleção e aquisição até a recuperação da informação ou de documentos nas unidades de documentação[3]. Tem por objetivo teórico estabelecer mecanismos para expressar de maneira o mais fiel possível, a representação dos elementos que pertencem ao conteúdo de um documento- seja ele real ou eletrônico- a fim de que o mesmo possa ser recuperado posteriormente. Neste estudo é apresentado o estado da arte da indexação documentária, os seus fundamentos teóricos e a indexação manual[4].

Palavras-chave: *Representação do conhecimento registrado; Indexação documentária; Indexação manual*

Recebido em: 19/04/2001 - Aceito para publicação em: 29/05/2001.

Introdução

O desenvolvimento científico e tecnológico favorece o aumento da produção do conhecimento, de um lado, e a sua fragmentação de outro, em função do aparecimento de novos campos do saber. Essa fragmentação, não implica em uma individualização das ciências, muito menos da tecnologia, muito pelo contrário, ambas buscam apoio intra e/ou entre elas e em outros domínios do conhecimento, a fim de serem melhor compreendidas nesse novo paradigma da sociedade dita da informação (*Information society*), do conhecimento (*Knowledge society*) ou do *aprendizado* (*Learning society*).

Neste contexto interdisciplinar, observamos que a produção do saber se apresenta formalmente estruturada sob várias formas, como por exemplo, sob a forma impressa (livros, periódicos, folhetos, patentes, relatórios técnicos, normas técnicas

[1] Artigo baseado na tese de doutorado: La representation des connaissances dans le contexte de la documentation tecnique: proposition d'um modèle d'indexation.

[2] Doutora em Ciências da Informação e da Comunicação, Institut de Communication et des Médias-Université Stendhal Grenoble 3-França. Professora do Departamento de Ciências da Informação, UFC–vbentes@ufc.br- Fone: (xx 85) 281 71 31-Fax: (xx 85) 243 41 40.

[3] Em todo o decorrer deste trabalho considera-se Unidades de documentação como sinônimo de bibliotecas, centros de documentação e outros do gênero. Portanto, trata-se de um espaço real ou virtual constituído por um acervo documentário independente do seu suporte e de sua forma.

[4] A complementação desse artigo trata da indexação semi-automática e automática e foi apresentada preliminarmente no X SNBU realizado em Fortaleza, em outubro de 1998.

etc.), sob a forma de imagem (fotografias, desenhos etc.), sob a forma de multimídia (combinação de textos, de imagens, de sons e outros dados). Além destas, existem ainda os conhecimentos tácitos, os materializados nos próprios produtos, como é o caso das máquinas e equipamentos, e também aqueles apresentados informalmente sob a forma oral em eventos: feiras, exposições, congressos, seminários.

Se a forma de apresentação do conhecimento mudou, seu suporte de estocagem também mudou e, hoje, encontramos os conhecimentos registrados tanto em suportes tradicionais como o papel, quanto em suportes eletrônicos, óticos e magnéticos.

Esse aumento do conhecimento se traduz pela disponibilidade de uma quantidade enorme de informação, recurso que está sendo considerado como o capital mais importante de nossa sociedade, posto que ele representa um alto valor agregado para o setor produtivo industrial e empresarial. Como a sociedade, a informação tem um papel fundamental para alavancar a ciência e a tecnologia, ela tem igualmente um lugar importante na vida cotidiana dos cidadãos, que precisam estar bem informados para que possam exercer dignamente a sua cidadania.

Foi a partir dessas observações que produzimos este artigo, tecendo comentários, inicialmente, a respeito do acesso à informação na sociedade do conhecimento, para, em seguida, trabalharmos a indexação em seu estado da arte e, finalmente, a indexação manual como forma de representação do conhecimento registrado visando a sua posterior recuperação.

Acesso à informação na sociedade da informação

A enorme quantidade de informação que hoje está disponível favorece a que se tenha a ilusão de que:

> *nunca estivemos tão informados, o que não quer dizer que sabemos tratar e integrar os dados que literalmente nos submergem. Muita informação mata talvez a informação, suscita evasões imaginárias, e recusa de saberes, e se choca de toda maneira ao 'segredo informacional' de cada um (um organismo só utiliza uma ínfima parte dos sinais que perpassam pelo seu meio ambiente* (BOUGNOUX, 1993, p.11).

Corroborando, Pierre LÉVY (1997, p.17), afirma que:

> *nós dispomos de poucas ferramentas para filtrar a informação pertinente, para fazer comparações de acordo com os significados e as necessidades sempre subjetivas, para nos deslocarmos no fluxo informacional* (LÉVY, 1997, p.24).

Essas duas reflexões mostram claramente o paradoxo dessa sociedade, a qual pode ser caracterizada, entre outras, como:

a) uma sociedade grande produtora e consumidora de informações sendo que, portanto, a sua matéria-prima é a informação e o conhecimento;

b) uma sociedade que, mesmo sendo produtora e consumidora de uma quantidade enorme de informações, se depara com inúmeras dificuldades para acessar esta avalanche de informações que nos submerge quotidianamente;

c) uma sociedade produtora e utilizadora das ferramentas de tratamento, estocagem e recuperação da informação, propiciadas pelas Tecnologias da Informação e da Comunicação (*TIC´s*);

d) uma sociedade na qual, se de um lado, as *TIC's* favorecem um alto alcance, onde predominam as lógicas das redes e a suposta flexibilidade, de outro, nos leva a conviver com a chamada info-exclusão e com inúmeras perdas;

e) uma sociedade na qual o ser humano necessita, ansiosamente, estar atualizado para poder acompanhar as transformações que se processam, quotidianamente, em uma velocidade infreável.

Dentre essas variáveis, as que concernem ao acesso e à recuperação de informações são infinitamente atingidas pelas outras, pois o ser humano – usuário - vive submerso pelas informações, necessita estar atualizado para não ficar a reboque das mudanças que se processam a todo o momento e se depara com as dificuldades para acessar tais informações, mesmo com a existência das ferramentas das TIC´s. Essas barreiras de acesso à informação são explicadas por KURAMOTO (1999), quando afirma que, para navegar sobre o *espaço do saber*[5] é preciso conhecer este espaço, saber utilizar as ferramentas das *TIC´s* e igualmente conhecer as estratégias de busca e recuperação de informação, ou ainda, ter a ajuda de um especialista neste domínio para que as necessidades de informação dos usuários possam ser atendidas de maneira eficaz e eficiente. Nesta perspectiva, torna-se imprescindível que saibamos o que fazer com as *TIC's* e quais as conseqüências de sua má utilização o que pode, certamente, comprometer tanto a tecnologia (produto/serviço) quanto seus usuários.

Além dessas variáveis, deve-se considerar, ainda, a ausência e/ou deficiência no processo de representação em nível de indexação. Esse fato é notório, tanto quando buscamos informações nas Unidades de documentação, ou ainda sobre a rede Internet que, quase sempre nos fornece como respostas um calhamaço de dados que, muitas vezes, não condizem com o que foi demandado. Por exemplo, um dia necessitava de uma informação sobre a minha conta no Banco do Brasil em Fortaleza. Estava na França e o acesso poderia ser feito pela Internet. Com a estratégia de busca Banco do Brasil & Fortaleza, procurei no Alta Vista e no Yahoo. Tive 193 respostas compostas pelas palavras banco, Brasil, Fortaleza e do ,mas nenhuma resposta com Banco do Brasil, o sintágma complexo. Problemas dessa natureza são enfrentados por inúmeros usuários da rede e certamente podem ter sido decorrentes de uma indexação baseada em unitermos, que considera cada palavra individualmente, e não os grupos sintagmáticos, daí, o fornecimento quase sempre de respostas insatisfatórias.

Este exemplo mostra, explicitamente, a necessidade de se investir no tratamento técnico dos recursos informacionais, assim como na sua organização, de maneira mais conveniente, visando a racionalização de sua estocagem e, naturalmente, a busca e a recuperação de informação de maneira eficaz e eficiente porque, como afirma CHAUMIER (1990, p.278), *de nada adianta arquivar um documento que não saberemos encontrar porque ele não foi indexado* ou, ainda, porque ele foi indexado de maneira incorreta.

Diante disso, perguntamos, o que significa, exatamente, a atividade de indexação? Como ela deve ser efetuada? Quais são os problemas enfrentados para a sua execução?

---

[5]A expressão espaço do saber foi proposta por Pierry LÉVY (1997).

Indexação: uma forma de representação do conhecimento

Para se trabalhar a indexação, inicialmente é preciso considerar que este conceito é utilizado em vários domínios do conhecimento, como na economia, na demografia, no comércio e nas ciências da informação. Esse último, é o que nos interessa.

Seguindo o raciocínio do professor Jean-Claude GARDIN (1974), consideramos a *indexação documentária* como um conjunto de atividades que consiste em identificar, nos documentos, os seus *Traços descritivos* (TD´s) ou macro-proposições e, em seguida, extrair os elementos/descritores (sintagmas) indicadores do seu conteúdo, visando à sua recuperação posterior. Esses descritores vão se constituir na representação dos elementos indicadores do conteúdo do documento e não a sua representação, pois esta só pode ser pelo próprio documento.

A representação do conhecimento registrado, tendo em vista a indexação de documentos, pode ser realizada tomando-se por base os conceitos, as palavras-chave/unitermos ou, ainda, em uma visão mais moderna, os sintagmas nominais (proposta apresentada pelo grupo *SYDO*), ou frases (proposta de Alain F. Smeaton e Paraic Sheridan), ou ainda os sintagmas verbais (proposta de Geneviève Lallich e de Virginia Bentes Pinto). No caso dos conceitos e das palavras-chave, eles podem ser extraídos[6] do documento mesmo ou ainda atribuídos a partir de outras fontes, como por exemplo as Linguagens documentárias (LD´s)[7]. Em contrapartida, os sintagmas ou as frases só podem ser extraídos do próprio documento.

A primeira forma de indexar nos parece representativa dos descritores constitutivos dos documentos, ou seja, do conjunto de suas unidades lexicais. Aqui é desmontado o discurso do autor, onde as palavras tinham um sentido em função do contexto ditado por seu criador; conseqüentemente elas eram ligadas ao mundo real do documento. Retiradas do seu contexto, tais palavras ou conceitos passam a significar apenas propriedades, portanto, seu sentido vai mudar, naturalmente (LE GUERN (1991).

Exemplo: *A representação do conhecimento registrado*

Na indexação por palavras-chave/unitermos, teremos o conjunto seguinte: *A, representação, do, conhecimento, registrado.*

Nessa maneira de indexar, as palavras são retiradas do contexto do lógico-semântico, onde elas tinham uma significação determinada por este contexto. Elevadas do mundo real, elas designam um conjunto de propriedades, seu sentido muda e se resumem a um conjunto de unidades lexicais. Ora, segundo Michel LE GUERN (1991, p.23)

> *a palavra da língua, contrariamente à sua ocorrência no discurso, não tem referência extra lingüistica(...) a relação signo/objeto, ao senso de Pierce, corresponde sumariamente à relação Saussuriana significante/significado, os significados fazem parte da estrutura da língua. Para que o descritor complete sua função, que é de colocar em relação um objeto*

[6] O primeiro caso é chamado por LANCASTER (1991) de indexação por extração e o segundo indexação por atribuição. PAIJMANS (1993) os chama de *assigned indexing* e *derivad indexing*.

[7] No decorrer deste trabalho, considera-se as linguagens de indexação (LD), como um conjunto de termos estruturados utilizados como tradutores dos elementos indicadores do conteúdo dos documentos visando a construção de índices para facilitar a recuperação da informação.

> *do mundo – uma entidade extra lingüistica – com o documento que trará as informações sobre este objeto, é preciso que o descritor seja um signo indiciário (...). As palavras da língua, enquanto que palavras da língua, significam apenas propriedades, nunca entidades: elas significam atributos, e não substâncias, até que elas sejam construídas no discurso. Quanto ao descritor, ele representa uma entidade, uma substância no sentido da filosofia de Aristóteles. Portanto, o descritor não pode ser considerado, a exemplo das palavras da língua, como sendo um símbolo sem referência.*

Ao contrário, se a indexação é realizada tendo como base os sintagmas ou as frases, os índices serão constituídos por passagens do texto portadoras de informação, neste caso, pode-se ter uma representação mínima do conteúdo do documento à medida em que esses grupos não são isolados do contexto no qual eles são inseridos (onde eles têm um valor referencial).

No exemplo apresentado anteriormente, se tomarmos como elementos representativos os sintagmas, iremos ter os seguintes:

Sintagma 1= A representação do conhecimento registrado

Sintagma 2 = o conhecimento registrado

Sintagma 3= o conhecimento

Nesse caso, podemos ter uma representação mínima do conteúdo veiculado no documento pois as suas unidades gramaticais não são retiradas do contexto, onde tinham um valor referencial. Além do mais, é preciso levar em consideração as características destes sintagmas, pois eles poderão ser portadores, seja de um conjunto de informações, seja de um simples fragmento.

A representação em nível de indexação documentaria perpassa, ao menos, por três etapas:

a) análise conceitual
b) tradução
c) controle de qualidade

Quanto à maneira de indexar, ela pode ser feita através de:

a) uma análise manual, chamada igualmente intelectual, que é feita pelos humanos;

b) uma análise mecânica, feita pelas ferramentas da informática;

c) uma análise que combina as duas: humana e mecânica. Esta ultima é chamada indexação semi-automática ou assistida pelo computador, sendo realizada da seguinte maneira: inicialmente, o sistema faz uma indexação automática dos documentos levando em conta as ocorrências das palavras mais freqüentes no texto. Em um segundo momento, o indexador humano refina a lista dos descritores propostos pelo sistema fazendo os ajustes e/ou complementações necessárias.

Com relação à dimensão da indexação, ela foi proposta pelo professor F.W. LANCASTER (1979), e trata-se da extensão com a qual um documento pode ser indexado, ou seja, a exaustividade (*exhaustivity*), e a especificidade (*specificity*). A indexação exaustiva procura extrair do documento o maior número de conceitos de forma a cobrir o seu conteúdo da maneira mais completa possível. É certo que esta

maneira de indexar oferece a oportunidade de acesso a um grande número de conceitos, mas, ao mesmo tempo, ela pode ser responsável pelo ruído[8] durante a recuperação da informação. Segundo SOERGEL(1994), a exaustividade pode ser vista sob dois aspectos: a exaustividade de pontos de vista e a exaustividade de importância. A primeira assegura que as facetas ou os pontos de vista considerados úteis para a representação proposta pelas LD's serão disponíveis para a recuperação da informação. A segunda determina o nível de importância dos descritores propostos pelas regras de indexação. Outro fator observado nesta maneira de indexar diz respeito ao seu custo-eficácia, pois, quanto mais exaustiva a indexação, naturalmente, maior será o seu custo, mas, ao se pensar na recepção desse produto, o retorno, certamente, será positivo à medida que o usuário poderá ter outras possibilidades de recuperação.

A indexação específica, como o nome o diz, leva em consideração os conceitos específicos em função dos temas tratados no documento. Esta maneira de indexar diz respeito à profundidade com a qual o conteúdo de um documento é tratado. Se de uma parte ela favorece a precisão, de outra, contribui para aumentar o silêncio[9] na recuperação da informação, pois é levado em consideração apenas o conteúdo principal do documento, deixando de fora outros assuntos tratados, mesmo que de maneira não elementar, e que poderiam responder às necessidades de quem está buscando informações.

Um outro aspecto que deve-se levar em conta na atividade de indexação refere-se à definição das Unidades de informação[10] , que poderão entrar na construção do índice. Em outras palavras, que fragmentos do texto devem ser levados em consideração: palavras? conceitos? sintágmas?

A maneira de indexar depende, naturalmente, do tipo de documento a indexar. Se tomarmos como exemplo uma monografia, como devemos indexá-la? Analisando o conteúdo predominante no documento, ou, de maneira mais fina, levando em conta, por exemplo, os capítulos, os parágrafos e as seções? No que concerne aos periódicos e aos anais de eventos, a indexação deve se apoiar sobre os artigos, sobre as conferências e comunicações ou a partir de seus títulos? Tratando-se da indexação de documentos técnicos - manuais técnicos, patentes, normas técnicas, bulas de medicamentos, laudos médicos etc., a indexação deve ser conduzida levando-se em conta os títulos, os capítulos, as passagens...?

Os estudos e experiências mostram que ainda não temos uma resposta precisa a estas interrogações. No entanto, o que se observa é que a indexação de monografias, de periódicos e de anais de eventos pode ser realizada tanto de maneira específica - levando-se em conta a estrutura lógica das monografias, dos artigos de periódicos e das comunicações em eventos - quanto de maneira bem geral, tomando-se como referência os títulos. Com relação à indexação de documentos técnicos, ela

---

[8]Consideramos como ruído, o excesso de documentos propostos pelas bases de dados para responder a uma demanda mas, na realidade, não respondem ao assunto demandado. Ele é medido pela relação entre o número de documentos não pertinentes fornecidos como respostas e o número total de documentos propostos: R=dn/d, onde dn=documentos não pertinentes, d=documentos da base.

[9]Neste trabalho, o silêncio corresponde à ausência de documentos que responderiam às necessidades dos usuários, mas, na realidade, não foram encontrados, mesmo que façam parte da coleção. A taxa de silêncio corresponde à relação entre o número de documentos pertinentes encontrados e o número total de documentos da base= Sl=dp/d, onde dp=documentos pertinentes, d=documentos da base.

[10]Consideramos como Unidades de informação os fragmentos de textos reconhecidos como possíveis elementos representantes de seu conteúdo.

deve ser conduzida no sentido da especificidade, pois os usuários deste tipo de documentos demandam uma informação pontual para responder às suas necessidades que são bem específicas. Assim, talvez, a solução seja indexar estes documentos levando-se em conta as suas estruturas lógicas, pois, normalmente, elas refletem com mais clareza o conteúdo tratado no documento. Desta forma, a indexação pode ser realizada a partir dos capítulos, seções, parágrafos, passagens etc.

No que concerne a indexação de documentos audiovisuais, cujas características são o conteúdo (informação), a mídia vetor deste conteúdo é o suporte de estocagem. Isso nos traz problemas pois, neste caso, a indexação demanda muito mais detalhes e muito mais informações do gênero: quem? o quê? como? onde? quando... ? A maneira de indexar estes documentos coloca em jogo além das informações visuais, outros tipos de informações percebidas por outros órgãos sensoriais desde que o sujeito conheça o conteúdo (CHELLAPPA, 1995). Segundo a professora Johanna W. SMIT (1989) as dificuldades para a indexação de documentos audiovisuais resulta da tentativa de passagem da denotação (o que o documento mostra) para a conotação (o que é percebido pelo indexador). Assim, como deve ser feita a indexação destes documentos? considerar os objetos representados e suas formas? a percepção visual ? a cena? o acontecimento?

Estas considerações mostram que não existe uma regra única para se elaborar uma representação em nível de indexação; a maneira segundo a qual um documento terá seu conteúdo representado deverá ser estabelecida pela política de indexação definida pelas Unidades de documentação. Esta política deve ser definida em função dos objetivos e da missão destes organismos, em função do perfil de seus clientes, e deverá estar contida em um manual, de maneira que os indexadores possam tomar conhecimento das regras estabelecidas e possam segui-las. Essa decisão poderá contribuir para reduzir, de certa forma, a subjetividade suscitada por esta atividade.

Indexação manual

A indexação manual, chamada igualmente intelectual ou humana, como o próprio nome o diz, é realizada pelos humanos, sejam eles bibliotecários ou especialistas do(s) domínio(s) no qual(is) essa atividade está sendo realizada. Esse tipo de indexação baseia-se, sobretudo, no julgamento, normalmente intuitivo, dos indexadores, em função do texto e do interesse para a sua comunidade de usuários.

Para realizar essa indexação, é preciso, inicialmente, analisar o conteúdo do documento, lendo-o não do inicio ao fim, mas por partes, ou seja, lendo suas estruturas lógicas. Por exemplo, a introdução, os capítulos, as seções, os parágrafos, a conclusão e outras passagens consideradas importantes. Essa análise pode ser estabelecida partindo-se das estruturas fornecidas pelos autores ou pelos editores de documentos ou ainda por uma segmentação proposta pelo indexador. Em resumo, ela comporta a leitura de documentos, a compreensão de seu conteúdo, a identificação e a seleção de conceitos para representar os elementos indicativos deste conteúdo. Segundo FIDEL (1994), se faz necessário considerar neste contexto dois aspectos: o documento propriamente, em outras palavras, o seu lado objetivo, assim como a razão ou motivo pelo qual o documento poderá ser utilizado, neste caso, os aspectos subjetivos. Aqui, a indexação manual enfrenta um grande problema: a dificuldade de se escolher os

conceitos que podem melhor representar os elementos indicadores do conteúdo do documento e a subjetividade desta escolha, o que pode explicar os desacordos freqüentes entre os indexadores humanos.

Após esta análise, passa-se a uma segunda etapa, a chamada tradução, na qual os indexadores fazem uma comparação entre os conceitos pré-selecionados em línguagem natural, com os descritores das LD´s. Se esses conceitos coincidirem com os das LD´s, eles poderão ser escolhidos como representantes dos elementos que fazem parte do conteúdo do documento. Na prática, sabemos que, se os conceitos selecionados não coincidirem com os descritores das LD´s, os indexadores poderão adota-los. Essa decisão dependerá de seu conhecimento sobre o assunto, de seu conhecimento sobre o perfil dos usuários, da política de indexação adotada e, igualmente, de sua experiência no domínio da indexação.

A tradução é uma etapa bem complexa, à medida em que exige o seguimento de regras que foram definidas a *priori*, como por exemplo, os tesauros, as listas de autoridades etc. Essas regras são consideradas como uma faca de dois gumes, pois, se de um lado elas parecem assegurar a qualidade da indexação, no que diz respeito à desambigüização das palavras, a organização e normalização dos índices, de outro, elas podem ser responsáveis pelo *silêncio* ou pelo *ruído* no momento da recuperação da informação. Assim, segundo FIDEL as experiências dos indexadores têm mostrado que, quando da etapa de tradução, é necessário levar em consideração alguns questionamentos, tais como:

a) *as fontes dos termos de indexação*: em quais fontes de vocabulários de indexação os indexadores podem se apoiar para escolher os termos que vão compor os índices ? Existem regras que limitam o indexador aos descritores dos tesauros utilizados pelo sistema, e outras permitem que sejam utilizados os termos da língua natural?

b) *a precisão*: que grau de precisão o indexador pode utilizar para traduzir os conceitos em termos de indexação? Os termos selecionados para o índice devem ser tão precisos que substituam o conceito ou eles devem ter um sentido mais geral?

Ex.: Fruticultura tropical no semi-árido: manga e caju
Este documento deve ser indexado por:
Frutas cítricas?
Manga?
Caju?

c) *o peso*: o peso relativo dos conceitos de um documento pode ser definido pelo indexador? No exemplo anterior, qual conceito terá o peso maior?

d) *a fidelidade*: em que medida a tradução deve ser fiel? Como ser fiel em uma tradução quando o conceito não tem um descritor correspondente? O indexador poderá usar os termos aproximados?

Ex.: Ciência→ Científico

e) a *linguagem do usuário*: o indexador pode designar os termos de um índice em uma linguagem mais próxima da do usuário? Por exemplo, através dos seus perfis é possível estabelecer regras que poderão guia-los na escolha dos termos de indexação mais adequados?

Ex.: Dor de cabeça→Cefaléia

Essas interrogações são de fundamental importância quando do estabelecimento das políticas de indexação a serem adotadas pelas unidades de documentação. As soluções das questões *a, b* e *e* são mais fáceis de serem resolvidas, porque elas são ligadas à tomada de decisão operacional. Ao contrário, as soluções das questões *c* e *d* são mais difíceis de resolver pois a definição do peso dos conceitos pertencentes a um documento implica em um processo subjetivo, portanto, difícil de se colocar em prática. Com relação à fidelidade da tradução dos conceitos que não possuem descritores equivalentes, talvez a solução seja conservar os conceitos, pois os descritores considerados próximos certamente que não possuem o mesmo sentido.

Além desses entraves, a indexação manual apresenta outras barreiras, entre as quais destacamos:

a) custo elevado; pois exige pessoal especializado e demanda muito tempo. Segundo BERTRAND (1994), o tempo médio de indexação de um documento é de aproximadamente 30 minutos, podendo variar de 10 a 45 minutos, segundo as dificuldades encontradas com relação à compreensão do conteúdo do documento;

b) fraca coerência intra e entre os indexadores; a coerência na identificação dos conceitos e na escolha dos termos de indexação, pelo mesmo indexador em momentos diferentes de indexação, ou por outros indexadores eqüivale a aproximadamente 30% (CLEVELAND, 1977);

c) a dificuldade para se escolher grande quantidade de conceitos; normalmente, no processo de indexação manual, a escolha dos conceitos não ultrapassa 5 (cinco). Assim, muitos assuntos tratados em um documento poderão não ser percebidos pelos indexadores, o que provocará aumento do silêncio no momento da recuperação da informação.

Na indexação manual, observamos que, se de um lado, o indexador tem um grande poder de decisão na definição dos conceitos que representarão os elementos indicadores do conteúdo do documento, de outro lado, encontramos vários inconvenientes, os quais causam problemas que são difíceis de administrar por que :

⇒ nem sempre o indexador é especialista no domínio do(s) assunto(s) que ele indexa;

⇒ nem sempre o indexador é especialista no domínio da indexação ;

⇒ existem novos domínios do conhecimento pouco cobertos ou ainda não cobertos pelas LD's especializadas;

⇒ as inovações terminológicas que se verificam em vários domínios do conhecimento exigem uma constante atualização tanto por parte das LD's, como pelos indexadores. Contudo, o que se verifica é que, na realidade, nem sempre os indexadores são treinados e/ou reciclados neste sentido, e muito menos as LD's são atualizadas.

Um outro aspecto a ser levado em consideração é que, embora a atividade de indexação manual pareça objetiva e neutra, na realidade isto não é verdade, pois ela consiste em um trabalho de síntese e, portanto, tem uma forte carga cognitiva, à medida em que demanda uma compreensão do conteúdo do documento, a extração dos elementos correspondentes a este conteúdo, a representação destes elementos e

a sua organização em forma de índices. Sendo naturalmente uma atividade subjetiva, ela é influenciada pelos conhecimentos sobre o domínio do documento, pelas experiências do indexador, pelo conhecimento da atividade de indexação, igualmente pelo contexto onde se realiza a indexação, entre outros. Assim, ela não poderá ser desprovida da neutralidade, mesmo que seja desejável.

Outra observação diz respeito à indexação manual: apesar da evolução dos processos de indexação semi-automática e automática, elas não excluíram esta prática. Assim, mesmo considerada como obsoleta para uns, a indexação manual ainda é utilizada tanto nos países do chamado primeiro mundo como nos do terceiro. Ora, se este tipo de indexação ainda é utilizada nos países grande produtores e consumidores de recursos informáticos isso ocorre, principalmente, por dois motivos:

⇒ porque a indexação semi-automática e automática não oferecem respostas totalmente satisfatórias no momento da recuperação da informação;

⇒ porque os sistemas de indexação automática ainda não atingem 100 % das unidades de documentação desses países.

Considerações finais

Nossa problemática neste trabalho foi a de apresentar o estado da arte da indexação, mostrando o seu conceito, assim como a sua prática, dando ênfase maior à indexação manual.

Observa-se que, graças à interdisciplinaridade da ciência da informação, notadamente com a informática, a estatística, a lingüística, e a psicologia cognitiva, o campo da indexação vem evoluindo bastante. Essa evolução é reconhecida através da literatura que mostra que, inicialmente, as experiências foram calcadas em uma prática manual simples e de forma intuitiva, cujo objetivo era o de fornecer um conjunto de palavras que fossem capazes de oferecer algumas pistas para o usuário encontrar o documento que necessitava.

Salientamos ainda que a indexação é uma atividade que desconstrói o discurso montado pelo autor do documento, à medida que faz recortes neste discurso. Assim, ela permite passar de um documento constituído (um documento primário) à sua re(constituição) em um novo documento - índice (um documento secundário) -, o qual é formado, não pela representação do documento inicial, mas pela representação dos elementos indicadores do seu conteúdo e que se constituirão na chave de acesso para a recuperação da informação. Além disso, a indexação coloca em cena três atores: o autor do documento, o indexador e o usuário. Consequentemente, para cada um a noção de pertinência informacional será percebida diferentemente segundo as suas experiências, os interesses de cada um no momento da produção ou de leitura do texto. Assim, a cobertura dos conceitos escolhidos para representar os elementos do conteúdo do documento não pode chegar a 100% pois esta indexação, nem sempre, é feita de maneira exaustiva, portanto, certas partes do documento não são levadas em consideração, o que provoca a perda de certas informações, e contribui para aumentar o silencio. No que concerne ao indexador e ao usuário, é desejável que a taxa de cobertura dos termos seja ótima, pois isto influenciará a pertinência das respostas fornecidas no momento da recuperação da informação.

Assim, qualquer que seja o método de indexação utilizado, manual, semi-automático e automático, a indexação, através dos componentes do índice, deverá permitir aos clientes o acesso ao documento que contém a informação que ele necessita. Seu resultado se constituirá em um dispositivo chave entre o documento primário a ser lido e compreendido, e um documento secundário (índice) a ser construído, de maneira tal que a representação dos elementos indicadores do conteúdo do primeiro sejam encontrados no segundo documento de maneira a mais completa e fiel possível. Pois é este documento índice que, durante a busca de informação, oferecerá *pistas* para que o usuário possa decidir, sem ver o documento primário, se ele irá considerá-lo ou não como possível para responder à sua necessidade. Portanto, a atividade de indexação que visa a representação dos elementos do conteúdo de documentos deverá ser calcada em dois objetivos fundamentais:

⇒ objetivo teórico: estabelecer os mecanismos para a elaboração dos índices;
⇒ objetivo operacional: possibilitar a busca e a recuperação da informação.

Para finalizar, lembramos que a indexação documentária é uma atividade que pode ser aplicada aos documentos textuais, visuais, sonoros, pictóricos, multimídia etc.

Document indexing: a form of representation of the registered knowledge
The domain of Information Science that has developed the most in the last 30 years is document indexing. This evolution is strictly linked to the changes of paradigms that happen in our society, independent of the domain of the knowledge. This article is concerned with the changes of indexing. It is the first part of a sequence of two papers about this topic.

Key Words: *Representation of the registered knowledge; Document Indexing; Manual indexing; Information Society*

Referências bibliográficas

PINTO, V. B. *La representation des connaissances dans le contexte de la documentation technique:* proposition d'un modèle d'indexation. Grenoble, 1999. Thèse de doctorat (Université Stendhal Grenoble-3).
BERTRAND, A. *Comprehension et categorisation dans une activité complexe:* indexation de documents scientifiques. Toulouse, 1993. Thèse de doctorat (Université de Toulouse).
BOUGNOUX, D. S*ciences de l'information et de la communication.* Paris: Larousse, 1993
CHAUMIER, J. L'indexation documentaire; de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique. *Documentaliste*, v. 27, n. 6., p. 275-284, nov./dec. 1990.
CHELLAPPA, R. Human and machine recognition of faces: a survey. *Proc. of. the IEEE*, v. 83, n. 5, May 1995
FIDEL, R. User-centered indexing. *JASIS,* v.45, n.8, p.572-576, 1994.
GARDIN, J.C. *Les analyses des discours*. Neuchatel: Delachaux et Nestlé, 1974.
KURAMOTO, H. *Proposition d'un système de recherche d'information assisté par ordinateur, avec application à la langue portugaise*. Lyon, 1999. Thèse de Doctorat (l'Université Lumiére).
LALLICH BOIDIN, G. *Analyse syntaxique automatique du fraçais application à l'indexation automatique*. Grenoble, 1986. Thèse de doctorat (Université des Sciences Sociales de Grenoble).
LANCASTER, F.W. I*ndexing and abstracting in theory and practice.* London: Library Association, 1991.

LANCASTER, F.W. I*nformation retrieval system characteristics, testing and evolution.* New York: J. Wiley, 1979. 381p. Cap.1, p. 1-14.

LE GUERN, M. Un analyseur morpho-syntaxique pour l'indexation automatique. *Le français moderne*. v. 59, n. 1, p. 22-35, 1991.

LÉVY, P. *L'intelligence collective*: pour une anthropologie du cyperspace. Paris: Découverte/ Poche, 1997. p. 21- 24.

SMEATON, A. F.; SHERIDAN, P. Using morpho-syntaxique language analysis in phrase matching. RIAO9: *Recherche d'information assisté par ordinateur*. Barcelona, 1991. v. 1, p. 414-430.

SOERGEL, D. Indexing and retrieval performance: the logiciel evidence. *JASIS*, v. 45, n. 8, p. 589-599, 1994.