Fábio Mascarenhas e Silva

Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica

Análise da Plataforma Lattes

Tese de doutorado
Área de Concentração Cultura e Informação
Linha de Pesquisa Acesso à Informação
Orientadora: Profa. Drª. Johanna Wilhelmina Smit

São Paulo
2007

Fábio Mascarenhas e Silva

Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica

Análise da Plataforma Lattes

Tese apresentada à Escola de Comunicação e Artes da Universidade de São Paulo como exigência para obtenção do título de Doutor em Ciência da Informação.

Área de Concentração Cultura e Informação

Linha de Pesquisa Acesso à Informação

Orientadora: Profa Drª Johanna Wilhelmina Smit

São Paulo

2007

Autorizo:

[X] divulgação do texto completo em bases de dados especializadas.

[X] reprodução total ou parcial, por processos fotocopiadores, exclusivamente para fins acadêmicos e científicos.

Assinatura: _____

Data: _____

A opinião, em homens de valor, é simplesmente o conhecimento em formação.

(Milton)

A

Deus, **minha esposa** (Adri), **meus pais** (Glauben e Tarcísio), **meus irmãos** (Cyntia “Nininha” e Tarcísio “Pipi”), dedico este trabalho com todo o amor.

AGRADECIMENTOS

À Profa. **Johanna Smit** pelas valiosas contribuições, sempre objetivas e claras, que contribuíram significativamente para o desenvolvimento deste trabalho.

Aos professores **Nair Kobashi e Raimundo Santos** pelo constante incentivo na minha carreira acadêmica, bem como pelas observações no Exame de Qualificação que muito ajudaram para o encaminhamento desta pesquisa.

A **Maria de Nazaré Ablas** que conseguiu, diante de tantos compromissos, se dedicar com carinho à revisão deste documento.

A todos os familiares, sobretudo os meus sogros **Clemência** e **José Luís**, que me encorajaram e apoiaram em importantes momentos vivenciados nos período da realização deste trabalho.

Ao grande amigo **Carlos Corrêa,** o “Mestre”, por todo apoio desde a fase mais embrionária desta tese até a sua definitiva conclusão e depósito.

Aos colegas do Departamento de Ciência da Informação da UFPE por todo o apoio, sobretudo aqueles que contribuíram diretamente na elaboração desta tese: **Maria Cristina Oliveira**, **Marcos Galindo** e **Susana Schmidt**.

Aos amigos **Renato Silva, Marivalde Francelin**, e **Rogério Mugnaini** pelo estímulo durante o doutorado.

RESUMO

SILVA, F. M. e. **Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica**: Análise da Plataforma Lattes. 2007. 163 f. Tese (Doutorado em Ciência da Informação) – Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 2007.

Discussão, avaliação e apresentação de parâmetros para a organização da informação científica e tecnológica (ICT) brasileira em meio eletrônico, enfocando os problemas do acesso à informação em sistemas abertos, especificamente a Plataforma Lattes do Conselho Nacional de Pesquisa (CNPq). Para fundamentação teórico-conceitual da pesquisa fez-se um retrospecto da ICT brasileira a partir da evolução das suas políticas nacionais de Ciência e Tecnologia e, em seguida, analisaram-se criticamente os recursos relacionados à organização da informação. Um estudo exploratório é apresentado, desenvolvido a partir de currículos extraídos da Plataforma Lattes, para identificar se a natureza aberta do sistema compromete a consistência dos dados na recuperação da informação. A análise se deu em duas etapas: a primeira, a partir da lógica dos Arquivos Pessoais e, a segunda, observando-se as formas de preenchimento do sistema a partir de três categorias: campos com Autonomia Total, Autonomia Parcial, e Sem Autonomia. Conclui-se que há comprometimento da consistência na recuperação da informação em sistemas abertos. A partir da sistematização dos resultados, apresentam-se sugestões para aprimorar o sistema.

PALAVRAS-CHAVE: Sistemas abertos; Sistemas de Recuperação de Informação; Informação Científica e Tecnológica - Brasil; Plataforma Lattes; Organização da Informação Científica e Tecnológica.

ABSTRACT

SILVA, F. M. e. **Scientific and technological information organization in open systems**: Lattes database analysis. 2007. 163 f. Thesis (Doctoral in Information Science) - Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 2007.

Discussion, assessment and presentation of parameters for organization of Brazilian Scientific and Technological Information (STI) on electronic means, focusing on the problems of access to information in open systems, specifically the Lattes Data Base of the Conselho Nacional de Pesquisa (CNPq). For the theoretical and conceptual well-grounding of this research, a retrospect of Brazilian STI was carried out from the evolution of its national Science and Technology politics, and then the resources related to the organization of information were critically analyzed. An explanatory study is presented, developed from CVs taken from the Lattes Data Base in order to identify if the open nature of the system puts the consistency of data at risk when information is retrieved. This analysis was carried out in two steps: the first one was done based on the logic of Personal Files, and the second one by observing the ways the system is fulfilled within three categories: fields with Total Autonomy, with Partial Autonomy and with No Autonomy. We conclude that consistency is at a risk when information is retrieved in open systems. From systemization of results we present suggestions to improve on the system.

KEY-WORDS: Information organization; Open systems; Information Retrieval systems; Brazilian Scientific and Technological Information; Lattes Data Base.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE SIGLAS

BDTD - Biblioteca Digital de Teses e Dissertações

CAPES - Coordenadoria de Aperfeiçoamento do Ensino Superior

CONSCIENTIAS - Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior

C&T - Ciência e Tecnologia

CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico

CV Lattes – Currículo Vitae Lattes

DTD - Document Type Definition

DGP - Diretório dos Grupos de Pesquisa no Brasil

DOD - Departament of Defense

FID - Federation International de Documentation

HTML – HiperText Markup Language

IBBD - Instituto Brasileiro de Biblioteconomia e Documentação

IBICT - Instituto Brasileiro de Informação em Ciência e Tecnologia

ICT - Informação Científica e Tecnológica

INPI - Instituto Nacional de Propriedade Intelectual

LD - Linguagens Documentárias

LMPL - Linguagem de Marcação da Plataforma Lattes

MCT - Ministério da Ciência e Tecnologia

MIT - Massachusetts Institute of Technology

NASA - National Aeronautics and Space Administration

NIT - Núcleos de Informação Tecnológica

NLM - National Library of Medicine

OMS - Organização Mundial da Saúde

OPAS - Organização Pan-Americana da Saúde

OWL - Web Ontology Language

PADCT - Programa de Apoio ao Desenvolvimento Científico e Tecnológico

PBDCT - Plano Básico de Desenvolvimento Científico e Tecnológico

P&D – Pesquisa & Desenvolvimento

PETROBRÁS – Petróleo Brasileiro S/A

PL - Plataforma Lattes

PND - Plano Nacional de Desenvolvimento

RDF - Resource Description Framework

RI – Recuperação da Informação

SciELO - Biblioteca Científica Eletrônica Online

SEICT - Sistemas Estaduais de Informação Científica e Tecnológica

SICT - Sistemas de Informação Científica e Tecnológica

SISTEMA CVLAC - Sistema de Currículos Vitae Latino-Americano e do Caribe

SRI - Sistemas de Recuperação da Informação

SGML - Standard Generalized Markup Language

SNICT - Sistema Nacional de Informação Científica e Tecnológica

TIC – Tecnologias de Informação e Comunicação

UNESCO - Organização das Nações Unidas para a Educação, a Ciência e a Cultura

UNICAMP – Universidade de Campinas

XHTML – Extensible Hyper Text Markup Language

XML - Extensible Markup Language

WWW - World Wide Web

APRESENTAÇÃO

O objeto de análise desta pesquisa foi a Plataforma Lattes (**PL**) do Conselho Nacional de Pesquisa (CNPq) em que, mais especificamente, explorou-se o sistema de gestão de currículos denominado Currículos Lattes. A apresentação da **PL,** em seu endereço na Internet[1], evidencia que há entre a **PL** e o Currículo Lattes uma inter-relação que dificulta dissociar um elemento do outro. Desta forma, doravante, a **PL** e o Currículo Lattes serão entendidos neste trabalho como um objeto único, mencionando-se apenas o termo **PL.**

A pesquisa sobre a **PL** foi conduzida a partir de um referencial teórico relacionado à organização da informação em meio eletrônico. As leituras críticas dos fundamentos teóricos nesta temática partiram da abordagem maior desta pesquisa: investigar se há comprometimento da consistência dos dados nos sistemas abertos de informação. A motivação para investigar tal assunto surgiu da percepção da crescente tendência de os próprios pesquisadores alimentarem os dados nos Sistemas de Informação Científica e Tecnológica[2] (SICT) brasileiros.

Desta forma, levanta-se a hipótese de que a atual metodologia adotada para coletar e organizar a informação na **PL**, ainda que elaborada a partir de estruturas computacionais bem definidas, pautadas em ontologias e linguagens de marcação, seja insuficiente para proporcionar uma organização da informação consistente e confiável.

O objetivo geral da pesquisa foi, portanto, discutir, avaliar e propor sugestões à organização da Informação Científica e Tecnológica (ICT) brasileira em meio eletrônico caracterizada pela livre inserção de dados nos sistemas. Para alcançar esse objetivo foi necessário cumprir algumas etapas, quais sejam: traçar um retrospecto histórico da ICT brasileira, visando contextualizar a evolução das suas políticas até os dias atuais; analisar criticamente os recursos voltados à organização da informação, identificando as vantagens e desvantagens de suas respectivas adoções; desenvolver estudo exploratório em um SICT nacional, a **PL**, com o propósito de identificar se há comprometimento na consistência dos dados decorrentes da natureza

[1] http://lattes.cnpq.br/conheca/con_hist.htm

[2] Para nosso trabalho o termo "Sistema de ICT" refere-se a recursos informacionais (produtos e/ou serviços) eletrônicos desenvolvidos para servir à comunicação e fluxo da ICT.

aberta do sistema; relacionar os procedimentos de organização da informação utilizados pela **PL** com recursos tradicionalmente utilizados para o tratamento da informação, como os vocabulários controlados, a fim de propor melhorias a partir do uso conjunto entre os recursos tradicionais e as novas formas de tratamento da informação.

1 INTRODUÇÃO

O processo comunicacional da ICT[3] é visto como uma atividade inerente a ambientes de pesquisa, quase um hábito natural àqueles que se inserem neste contexto. Aceita-se que divulgar resultados de experimentos e pesquisas tornou-se um compromisso, se não social, ao menos profissional dos atores envolvidos no aprimoramento, inovação, ou refutação do conhecimento universal.

Nestes espaços, a comunicação entre os pares é caracterizada pela delimitação de códigos lingüísticos consensuais (e contextuais). Assim, grupos afins adotam terminologias que delimitam seus objetos de estudo e respectivas concepções dentro de um domínio de conhecimento, promovendo uma comunicação mais restrita intentando torná-la mais eficaz para quem dela faz uso.

Entretanto, como a comunicação da informação não está restrita a uma delimitação lingüística, houve a necessidade de se desenvolver produtos e serviços orientados ao fluxo dos estoques de informação produzidos pelos atores que compõem o contexto da Ciência e Tecnologia (C&T).

Essa orientação à ICT, mais perceptível a partir da segunda metade do século XX, contou com novas áreas de pesquisa como a Ciência da Informação, Computação, Comunicação, Lingüística, entre outras, e resultou no desenvolvimento e aperfeiçoamento de instrumentos e metodologias tecnicamente mais rigorosos, como as Linguagens Documentárias e as Bases de Dados.

Posteriormente, tal esforço contribuiu também para a evolução das tecnologias eletrônicas de representação e recuperação da informação, produzindo recursos utilizados amplamente nas buscas em redes eletrônicas de comunicação.

Os sistemas genéricos de busca[4], como o Google, apresentam deficiências para encontrar informações mais especializadas, porém tais

[3] O Conceito de ICT adotado neste trabalho fundamenta-se numa nova percepção da ICT que, entre outras características, incluí as informações "demandadas pelas interfaces da produção científico-tecnológica com o Estado e suas instâncias decisórias, no planejamento e gestão de C&T" (GONZALEZ DE GÓMEZ; CANONGIA, 2001, p.12).

[4] Realizam buscas por qualquer tipo de informação em servidores da Internet. Em alguns casos como o Google Acadêmico há uma restrição por servidores de Universidades, Editores Científicos, Bases de Dados.

sistemas não serão discutidos nesta pesquisa, pois este estudo limitar-se-á à ICT, caracterizada por sistemas de informação – idealmente - produzidos a partir de um planejamento prévio e com finalidades específicas.

Para Rowley (2002, p.131), o planejamento prévio de um sistema de informação deve levar em consideração o “Ciclo de vida dos Sistemas”, composto por seis etapas: *análise*, *projeto*, *implementação*, *evolução operacional*, *deterioração* e *substituição*. Nas etapas do Ciclo de Vida dos Sistemas insere-se o desenvolvimento propriamente dito, compreendendo a definição de objetivos e requisitos, a elaboração do projeto, a implementação e, por fim, a avaliação. Esses passos são comumente orientados por uma lógica racional (busca-se uma melhor relação entre o custo e benefícios dos sistemas tornando-os economicamente rentáveis) e objetivam uma melhor produção, organização e disseminação dos estoques de informação.

Estes estoques de informação, ainda que produzidos na intenção de fazer avançar o conhecimento, quando não utilizados, são meros acervos armazenados em bancos de dados, os quais, para cumprirem seu papel de transmissores do conhecimento, requerem uma efetiva comunicação com seus receptores (Barreto, 1994). A referida comunicação se dá a partir do tratamento da informação para fins de recuperação, e almeja dispor recursos que facilitem a intermediação entre as necessidades de busca dos usuários com as estruturas significantes contidas nos acervos, independente destes acervos estarem em meio impresso ou eletrônico.

Especificamente no meio eletrônico, tal tratamento tem provocado debates entre grupos distintos. Apenas para fins explicativos, tais grupos foram categorizados em três: os que defendem a organização da informação de forma manual, incluindo a participação humana na análise dos documentos; os que crêem em melhores resultados através de um tratamento totalmente automático dos estoques; e, aqueles que acreditam que a integração entre os dois recursos é a opção mais viável para uma nova realidade pautada em estoques híbridos (impressos e eletrônicos). Para esta pesquisa, a última forma é vista, atualmente, como a mais indicada para o tratamento da informação.

Assim, pesquisadores da área da informação buscaram soluções a fim de amenizar os problemas identificados no processo de Recuperação da Informação (RI). No entanto, parte-se do princípio que investigações científicas

atreladas a atividades como a produção, comunicação e uso da informação devem considerar o contexto no qual se pretende desenvolver novos recursos, pois essa é uma condição básica para lidar com um produto cultural dotado de significado.

1.1 PROBLEMA

Um dos pilares desta pesquisa é uma tendência que vem ganhando força: a facilidade de os próprios autores produzirem não apenas o conhecimento propriamente dito, mas também a representação desse conhecimento nos SICTs. Essa mudança, visível em recursos como os *arquivos abertos*, periódicos científicos, ou em grandes sistemas de ICT como a **PL** (a caracterização da **PL** encontra-se em Anexo), se por um lado facilita a disponibilização/acesso aos documentos, por outro pode dificultar a RI e o uso dos dados para produção de indicadores em C&T.

Acredita-se que as iniciativas voltadas à organização da informação que alcançaram resultados mais consistentes compreendem que a informação é um produto social, atrelado a sistemas de significação construídos por indivíduos e grupos, e que aqueles que planejaram tais iniciativas perceberam que a velocidade de processamento das Tecnologias da Informação e Comunicação (TICs) potencializaria as tarefas dos sistemas de informação, mas não solucionaria todos os problemas atinentes à organização da informação.

Nesse contexto, instrumentos para representação, organização e comunicação da informação foram desenvolvidos, tais como as **linguagens documentárias**, **vocabulários** e ***tesauros***, que, integrados a ferramentais informáticos, aperfeiçoaram os sistemas automatizados de RI. A importância desses instrumentos na mediação do processo de transferência da informação foi debatida por autores como Tálamo (1997), Cintra (2002), García Gutierrez & Lucas Fernández (1987), Hutchins (1978), Van Slype (1991), e Sowa (1984).

Autores como Buckland (1997) diferenciam um documento de seu conteúdo: essa distinção é essencial para perceber que o suporte do conhecimento não altera o conhecimento registrado, o que prevalece é a informação contida no documento. Este posicionamento é defendido por Alvarenga (2001) ao esclarecer que o conceito é um elemento invariável e

também por Campos (2002, 2004), que busca entender a melhor maneira de representar hiperdocumentos, ou ainda pelas análises de Biolchini (2001) sobre os vocabulários controlados e ontologias em bases de conhecimento.

Outras pesquisas defenderam que o tratamento da informação em meio eletrônico não dispensa o laborioso trabalho de contextualização do conhecimento. Tal intervenção é ainda necessária para, por um lado, constituir os estoques baseados muitas das vezes em premissas econômicas e políticas (SAYÃO, 1996), e por outro, reduzir diferenças culturais, regionais e também terminológicas. Por isso, Dias,E.W.(2001) pondera que há necessidade de harmonização entre o tratamento da informação de forma híbrida (automática e humana), sugestão essa cada vez mais aceita como um caminho promissor, principalmente para a construção de ontologias para uso na Internet.

Em outra investigação, Rada (1991) apresenta iniciativas híbridas tais como: o esforço da ONU para construir uma terminologia unificada visando classificar documentos das Ciências Sociais; a padronização das linguagens de indexação da *Armed Services Technical Informacion Agency* e Atomic Energy Commision; e o desenvolvimento de um sistema unificado de linguagem da área médica, por parte do National Library of Medicine. Destaca-se que esses experimentos alcançaram bons resultados em razão de os documentos, antes de serem digitalizados, terem sido anteriormente tratados, caracterizando uma política pré-definida voltada à organização da informação.

Não restam dúvidas sobre a importância em tornar acessíveis artigos, trabalhos, livros, relatórios, e outras produções técnico-científicas. Sabe-se que há um entusiasmo por parte da comunidade acadêmica quanto à criação e uso de recursos informacionais que disponibilizem a produção dos meios acadêmicos, mas a massificação de meios de divulgação científica, sobretudo através da World Wide Web (WWW), talvez proporcione resultados contrários aos desejados, ou seja, criem-se obstáculos para a utilização da ICT.

Em estudo de Bergman (2001), percebe-se o grande volume de documentos “invisíveis” aos sistemas de busca da Web. Apesar de o trabalho desse autor ter englobado conteúdos genéricos, é possível considerá-lo como um sinal da alta incidência de informação não recuperável por sistemas como o Google, Yahoo, Altavista, Excite, etc. Bergman cunhou duas expressões: “*Surface Web*” e “*Deep Web*”. A *Surface Web* representa a parcela da WWW

recuperável pelos sistemas de busca e a *Deep Web,* a parcela "invisível" da Web que não é recuperada pelos sistemas de busca mais utilizados e acima mencionados.

Numericamente, há uma diferença enorme entre as duas, pois enquanto a *Surface Web* contém 1.000.000.000 (um) bilhão de documentos, a *Deep Web* 500.000.000.000 (quinhentos bilhões). Além disso, a *Deep Web* é caracterizada por conteúdos mais especializados e menos genéricos e, de acordo com o autor, "o total de conteúdo de qualidade na *Deep Web* é de 1.000 a 2.000 vezes maior que na *Surface Web*" (BERGMAN, 2001)[5].

Ainda que requeira maior aprofundamento, o estudo de Bergman evidencia que há um considerável volume de informação oculta aos sistemas mais populares de busca utilizados para recuperar informações na Web. Percebe-se que não há uma relação direta entre a massificação dos meios de disponibilização da informação e as possibilidades de recuperar informação qualitativa. Infere-se que nem sempre haverá mais e/ou melhor informação recuperada se maior for a quantidade de informação disponibilizada.

Como o foco deste trabalho é a ICT, ressalta-se a preocupação de autores diante do fato de que a aceitação dos recursos eletrônicos para a produção/disponibilização da ICT é um processo ainda a ser assimilado pela comunidade científica. Na opinião de Capurro (2002), a América Latina necessita de ações para o desenvolvimento de uma cultura digital, que na visão do autor só surgirá se os latino-americanos a criarem por si próprios, para si próprios e para os outros. Esse autor defende que há muito a se conquistar além do domínio das tecnologias eletrônicas, pois soluções relacionadas às tecnologias digitais configuram somente parte do problema.

Iniciativas louváveis, por defenderem a socialização do conhecimento podem contribuir para promover efeitos contrários, ou seja, criar obstáculos para a busca da informação. Tais obstáculos podem resultar da convergência à disponibilização/acesso da informação e desatenção à organização/recuperação.

O desequilíbrio de esforços (ênfase ao acesso e pouca atenção à organização da informação) não acarretará prejuízos para o crescimento do

[5] Por serem documentos eletrônicos sem paginação, algumas citações transcritas indicarão apenas o ano da publicação.

movimento de livre acesso, pois o barateamento dos equipamentos de informática, a oferta crescente de sistemas gerenciadores de informação gratuitos e a facilidade de produção e reprodução de documentos eletrônicos serão aliados a curto, médio e longo prazos.

Se já existem problemas concretos com o volume atual dos acervos digitais, é mais que urgente ocupar-se com propostas voltadas à organização dos conteúdos disponíveis, pois os instrumentais técnicos para disponibilizar conteúdos já foram bem simplificados (e massificados).

Uma preocupação ainda maior diz respeito à facilidade e flexibilidade de os usuários inserirem, além dos documentos, as representações de suas produções científicas e técnicas, ou seja, seus metadados. Essas representações tanto dizem respeito às descrições físicas como temáticas dos documentos digitais.

Na representação descritiva há problemas devido à falta de padronização, porém soluções técnicas menos complexas podem criar mecanismos que direcionem e orientem os usuários a alimentarem os SICTs. No caso da descrição temática, exige-se a habilidade do pesquisador para descrever tematicamente seu trabalho, criar relações hierárquicas e associativas e ainda categorizar o conteúdo dentro de um domínio específico de conhecimento. Permanecem dúvidas se os atores da ICT nacional conseguem compreender a finalidade e os fundamentos que existem por trás dessas representações.

Há discussões na literatura brasileira da Ciência da Informação sobre o assunto. É o que se vê em recente publicação de Marcondes (2006), ou em um outro trabalho - com participação do mesmo autor - no qual são debatidas as novas formas de cooperação em ICT (MARCONDES e SAYÃO, 2002). Salienta-se que, nesse último, a referida cooperação condiz com recursos de interoperabilidade entre sistemas de informação, que dependem de coincidências sintáticas entre conteúdos, ou seja, são pré-definidas relações de equivalência entre campos e seus respectivos atributos.

Outro debate relacionado a esse assunto foi visto com Pacheco e Kern (2001). Estes, ao explorarem a **PL** do CNPq, analisaram a estrutura da linguagem de marcação da referida plataforma, buscando entender como os dados deste sistema estão estruturados descritivamente e semanticamente. A

partir do entendimento do sistema, os autores defenderam a criação/implantação de uma ontologia comum para sistemas de informação e conhecimento sobre a C&T nacional.

Na prática, a proposta dos autores almejou estabelecer mecanismos que garantissem maior uniformidade aos dados e, consequentemente, um maior nível de consistência nas relações entre eles. Na visão dos autores, essa uniformidade proporcionará, entre outras vantagens, maior confiabilidade nos indicadores de produção científica. Este enfoque é interessante, porém não esclareceu como podem ser estabelecidas relações semânticas confiáveis.

É visível que as gestões da ICT, em alguns países, são mais consolidadas que a brasileira e alcançaram um nível de organização e desenvolvimento diferenciados, baseados em procedimentos e práticas bem definidos, orientados por um processo de tratamento da informação pautado em maior rigor técnico, objetivando constituir estoques de informação a partir de um viés produtivista, gerenciado por uma racionalidade econômica.

Nos sistemas de informação bem organizados, se houver rigor na gestão dos estoques (o que não implica ausência de falhas) o uso da linguagem natural como meio para recuperar a informação se torna mais viável. Contudo, a viabilidade ocorre em razão de haver uma gestão contínua dos SICTs. Assim, almejar eficientes sistemas nacionais de ICT necessita, antes de implantar softwares, estabelecer princípios quanto à organização da informação.

Um fato que se torna cada vez mais visível é a oferta crescente de produtos e serviços informacionais produzidos com a intenção de facilitar ao máximo a comunicação da informação através de redes eletrônicas de informação, destacando-se a Internet. Sabe-se que a Internet é uma rede mundial de computadores e que seu recurso mais conhecido é a World Wide Web (WWW ou Web), formada por, entre outras coisas, uma enorme quantidade de documentos armazenados (e acessíveis) em servidores.

Enquanto as publicações na Web se limitavam a serviços comerciais gratuitos (páginas pessoais, blogs[6], fotologs[7], etc) sem compromisso formal, problemas quanto à qualidade, veracidade e propriedade intelectual dos

[6] Um weblog ou blog é uma página da Web com mensagens textuais (posts) organizadas cronologicamente. Estes posts se referem a inúmeros assuntos, mas refletem normalmente as opiniões pessoais daqueles que os mantém.
[7] Similar ao Blog, diferencia-se por dispor mais imagens que textos.

conteúdos moviam discussões no meio acadêmico. Contudo, novos serviços e produtos informacionais foram criados e aperfeiçoados para grupos especializados, incluindo aqueles do universo científico e tecnológico. Facilitou-se assim o processo de divulgação da ICT.

A percepção de Targino (2002) sobre o advento de novos recursos eletrônicos como meio para que atores da C&T publiquem mais facilmente é negativa. A autora critica severamente a inconsistência das informações e a complexidade de armazenamento e controle bibliográfico. Mesmo não explicitando a preocupação com aspectos da organização da informação, Targino (2002) demonstra-se atenta à ausência de controle na alimentação de estoques de ICT, e afirma que

> publicações eletrônicas que se propõem à atualização imediata de informações são disponibilizadas de forma irregular e descontínua, [...] em termos genéricos, os registros não passam por um filtro que garanta a qualidade dos dados. Prioriza-se o crescimento quantitativo da Rede, em detrimento dos aspectos qualitativos e dos seus impactos sociais, o que repercute no ciclo da informação e, por conseguinte, nos processos de comunicação científica.

Uma modalidade recente de comunicação científica eletrônica, são os já mencionados arquivos abertos, caracterizados pela facilidade de publicação pelo próprio autor. Nessa modalidade, os usuários têm autonomia para inserir documentos no sistema, descrevê-los e classificá-los, razão pela qual esse procedimento é também conhecido como auto-arquivamento.

Para Café e Lage (2002) o auto-arquivamento garante ao autor

> a visibilidade e acesso aos trabalhos de pesquisa desenvolvidos, aumentando as possibilidades de ser citado e conhecido amplamente. Além disso, minimiza radicalmente as barreiras impostas nos sistemas tradicionais de publicação.

Concorda-se que o auto-arquivamento simplifique o acesso aos documentos, porém defender o acesso irrestrito à ICT não pode excluir um ponto fundamental: a recuperação da informação.

À medida que haja crescimento dos recursos de auto-arquivamento, é provável que haja um proporcional aumento de inconsistências decorrentes da falta de controle na gestão dos estoques (principalmente na inserção dos registros). Assim, uma das motivações desta pesquisa foi investigar se há de fato comprometimento da consistência nos sistemas abertos de informação. Para fins de análise o estudo foi delimitado a um objeto da ICT brasileira: a **PL** do CNPq.

Apesar dos grandes avanços alcançados pela **PL** nos últimos anos, ainda é preciso aperfeiçoar as mediações deste sistema com os usuários, e imagina-se que uma alternativa seja adotar mecanismos de controle adequados aos princípios de organização da informação. A ausência de tais mecanismos sugere que não se previu, na etapa de planejamento da **PL,** que o preenchimento dos currículos seria feito por uma comunidade bastante heterogênea e nem sempre familiarizada com recursos de informação.

1.2 JUSTIFICATIVA

É necessário refletir sobre os aspectos de organização da ICT em meio eletrônico nacional, pois o conjunto de procedimentos necessários para o desenvolvimento, implantação e manutenção de qualquer sistema de informação requer um grau de conhecimento daqueles que alimentarão e/ou modificarão os registros do sistema, bem como o uso que deles será feito. Por registro, entenda-se cada novo documento inserido, que no caso da **PL** é um currículo de pesquisador.

Dispor a qualquer indivíduo da comunidade acadêmica as chances de alimentar um SICT pode resultar em situações de difícil possibilidade de reversão ou mesmo de irreversibilidade, provocadas pela ausência de controle na entrada dos dados.

1.3 HIPÓTESE

Partiu-se da hipótese de que a atual metodologia adotada para coleta e organização da informação na **PL**, ainda que elaborada a partir de estruturas computacionais bem definidas, pautadas em ontologias e linguagens de marcação, seja insuficiente para proporcionar uma organização da informação consistente e confiável. Tal problema compromete o processo de recuperação da informação, e também a geração e uso dos dados da **PL** para fins de gestão da C&T.

1.4 OBJETIVOS

Diante do exposto, estabeleceram-se para esta pesquisa os seguintes objetivos:

Objetivo Geral

Discutir, avaliar e propor sugestões à organização da Informação Científica e Tecnológica brasileira em meio eletrônico caracterizada pela livre inserção de dados nos sistemas, tomando por exemplo a **PL**.

Objetivos Específicos

- Traçar um retrospecto histórico da ICT brasileira, visando contextualizar a evolução das políticas até os dias atuais;

- analisar criticamente os recursos voltados à organização da informação e identificar as vantagens e desvantagens de suas respectivas adoções;

- desenvolver estudo exploratório na Plataforma Lattes, na condição de um SICT nacional, com o propósito de identificar se há comprometimento na consistência dos dados decorrentes da natureza aberta do sistema;

- relacionar os procedimentos de organização da informação utilizados pela **PL** com recursos tradicionalmente utilizados para o tratamento da informação, como os vocabulários controlados, a fim de propor melhorias**.**

1.5 METODOLOGIA DE ANÁLISE

O objeto teórico deste estudo é a organização da ICT em meio eletrônico em sistemas abertos, e o objeto da análise exploratória foi a **PL** do CNPq. A escolha pela **PL** é justificada pela importância atribuída pela comunidade científica brasileira a esse sistema e também pelo volume de dados que compõe o acervo do sistema[8].

Para uma análise mais bem fundamentada construiu-se um quadro referencial teórico baseado na bibliografia científica, maiormente da área da Ciência da Informação. A primeira temática explorada é a Informação Científica e Tecnológica, convergindo a uma retrospectiva histórica crítica da evolução da política da ICT nacional. Em seguida, foram abordados os conceitos concernentes à organização da informação que contribuíram para as análises feitas no objeto desta pesquisa.

No projeto original deste estudo previa-se que as análises seriam feitas a partir de dados extraídos em formato XML da **PL**. Com os currículos em formato XML seria viável formatar os dados e utilizá-los em aplicativos específicos para análises bibliométricas e ainda seria possível conduzir as análises a partir de amostragens suficientemente representativas em termos estatísticos.

Porém, por haver restrição de acesso à base de currículos (para extração[9]), seria necessário que o CNPq autorizasse o acesso ao sistema ou que enviasse os currículos já coletados. Infelizmente, a negociação com a referida instituição governamental não evoluiu e por tal razão, as análises foram feitas em duas etapas distintas:

Etapa 1 – Análise da PL a partir da lógica dos arquivos pessoais

Fez-se uma avaliação crítica dos currículos da **PL** a partir de uma percepção arquivística: a lógica dos arquivos pessoais. Para tanto confrontou-se a visão arquivística com o modelo do currículo que é gerado pela **PL**. Visando tornar a explicação mais clara, usou-se um currículo de pesquisador como exemplo, mantendo sua identificação no anonimato.

---

[8] Em agosto de 2007 a PL ultrapassou um milhão de currículos (http://www.cnpq.br/saladeimprensa/noticias/2007/0820c.htm).
[9] http://lattesextrator.cnpq.br/lattesextrator/index.jsp

Etapa 2 - Análises do preenchimento da Plataforma Lattes

Em razão dos detalhes desta etapa serem exaustivos, preferiu-se descrevê-los na própria seção de análise. Entretanto, antecipa-se que foram criadas três categorias para as formas de preenchimento dos campos: **Autonomia total** (o usuário tem a liberdade de cadastrar as palavras que desejar sem restrição); **Autonomia parcial** (campos inicialmente livres, porém, cada novo termo cadastrado pelo usuário é automaticamente armazenado no sistema, que vai criando uma lista de termos exclusiva do usuário); **Sem autonomia** (o sistema prevê opções que o usuário deve selecionar).

Ressalta-se que o estudo exploratório aqui proposto não persegue a exaustividade, mas a discussão de aspectos da PL que refletem na RI.

2 A INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA

Na história da Ciência e da Tecnologia há diversos elementos que contribuíram diretamente para o desenvolvimento dessas duas instituições sociais. Entre esses elementos, destaca-se a informação, que certamente participou de todas as etapas da consolidação daquilo que hoje é visto como científico e/ou tecnológico. Tal afirmativa é justificada pelo próprio papel da informação, que ao longo do tempo, vem preservando, disseminando e proporcionando a produção de novos conhecimentos.

Não seria arriscado afirmar que não haveria C&T sem a informação, pois ambas se valem, acima de tudo, das percepções e interações do homem com o mundo em que vive e, nessa vivência, o progresso se fez graças ao contínuo esforço intelectual, não de apenas um, mas de vários sujeitos, que perpetuaram seus saberes registrando e comunicando suas descobertas.

A informação produzida e utilizada no contexto da C&T é descrita na literatura como ICT ou Informação em Ciência e Tecnologia. No intuito de definir e caracterizar melhor a ICT, optou-se por expor, de forma isolada, os conceitos de Informação Científica e Informação Tecnológica.

A Informação Científica para Aguiar (1991), - fundamentado em relatório da Federation International de Documentation (FID) - é todo conhecimento produzido ou que tenha relação com resultados de pesquisas científicas. O autor (p.9) define, por outro lado, a Informação Tecnológica como “todo conhecimento de natureza técnica, econômica, mercadológica, gerencial, social, etc. que, por sua aplicação, favoreça o progresso na forma de aperfeiçoamento e inovação”.

Como a Ciência e a Tecnologia são mutuamente atreladas, principalmente em áreas como a Química, a Física, e a Matemática, as pesquisas científicas e o desenvolvimento e inovação de novos produtos têm, entre alguns insumos, o conjunto de informações produzidas por atores da área científica e tecnológica. Assim, a ICT é

> constituída de elementos simbólicos utilizados para comunicar o conhecimento científico e técnico, independente de seu caráter (numérico, textual, icônico, etc.), dos suportes materiais, da forma de apresentação. Refere-se tanto à substância ou conteúdo dos documentos quanto à sua existência material. Também se emprega este termo ICT para designar tanto a mensagem (conteúdo e forma) quanto sua comunicação (ação). Quando necessário, distingue-se entre informação bruta (fatos,

> conceitos, representações) e os documentos em que se acha registrada (UNISIST II citado por Aguiar 1991, p.8*).*

A definição do UNISIST II ainda é aceita, contudo, houve consideráveis mudanças no contexto no qual a ICT está inserida. Por essa razão, no trabalho coordenado pelo IBICT (GONZALEZ DE GÓMEZ; CANONGIA, 2001, p.12) a definição de ICT foi ampliada por considerar-se que essa modalidade de informação não se restringe a um conjunto de conhecimentos produzidos e utilizados por cientistas e tecnólogos sobre temas de suas respectivas áreas de atuação, mas a

> toda a informação que os cientistas e as organizações de P&D precisam para desenvolver suas atividades [...]; as demandadas pelas interfaces da produção científico-tecnológica com o Estado e suas instâncias decisórias, no planejamento e gestão de C&T; e finalmente, informações destinadas a ampliar a participação da cidadania e suas expressões organizadas nos processos de elaboração de políticas públicas.

Portanto, nesta pesquisa, o conceito de ICT inclui informações que servem de apoio à gestão da C&T e também de instrumento para que pesquisadores possam compartilhar e conhecer suas produções. E a **PL**, como sistema de informação curricular, serve tanto como recurso informacional para instituições como para toda a comunidade científica brasileira (pesquisadores, estudantes, gestores, profissionais e demais atores do sistema nacional de Ciência, Tecnologia).

2.1 A COMUNICAÇÃO DA INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA

Para Le Coadic (2004), a informação insere-se num ciclo composto por três fases: a construção, a comunicação e o uso. Elas se sucedem e são interdependentes. Tal modelização é simplificada e descreve genericamente os processos da informação. No âmbito das discussões a respeito da comunicação da informação há um assunto restrito, porém bem difundido e consolidado: a comunicação científica, que neste trabalho será chamado de comunicação da ICT.

O processo de comunicação da ICT, segundo Meadows (1999), originou-se na Grécia Antiga, e a Academia foi o primeiro ambiente destinado à disseminação (oral) das reflexões sobre o mundo. A tradição escrita no universo acadêmico iniciou-se com os discursos de Aristóteles registrados em

manuscritos. O hábito da escrita como registro de reflexões expandiu-se para a cultura árabe e, posteriormente, para a Europa Ocidental.

Em seguida, algumas inovações técnicas permitiram que o conhecimento fosse compartilhado de forma mais eficiente, não restando dúvidas de que a introdução da imprensa na Europa, no século XV, contribuiu consideravelmente para o crescimento das publicações no mundo.

Mas, foi a partir da criação, em 1662, da ***Royal Society*** de Londres que a Ciência sistematizou a preocupação com a comunicação de suas descobertas. Entre as razões que favoreceram a publicação dos primeiros periódicos científicos, destaca-se o interesse dos editores em aumentar os lucros a partir da melhoria do processo da comunicação científica. A intenção dos editores era despertar ainda mais o interesse por novidades em seu público potencial: os cientistas.

Os avanços da comunicação da ICT modificaram a maneira de compartilhamento e de contribuições entre os pares. A esse respeito, Mathias (1972) explica que o avanço do conhecimento científico foi um fenômeno europeu, assim como foi também na Europa - principalmente na Inglaterra e França - que se iniciou o vínculo entre a ciência e a técnica visando à aplicação dos resultados na indústria e agricultura.

Outra interessante abordagem foi feita por Wersig (1993), que analisou as mudanças ocorridas no papel do conhecimento, dentre as quais se destaca a *fragmentação do conhecimento*. O autor cita três razões que proporcionaram a expansão do conhecimento: o grande volume de informações, a criação de padronizações próprias por cada área de conhecimento e o pluralismo de opiniões e visões de mundo.

A tendência é que o pluralismo continue a crescer, sendo sustentado pela multiplicidade das tecnologias da informação. Durante o predomínio da palavra falada e impressa, havia uma maior limitação dos mecanismos técnicos disseminadores do conhecimento. Hoje, o incremento dos recursos eletrônicos e a diversificação das organizações e mídias de apresentação contribuem para a diversidade e crescimento de produtos informacionais especializados como bases de dados, livros e periódicos.

As publicações técnicas e científicas começaram a vivenciar um crescimento vertiginoso. Isso foi percebido por Weisman (1972) há mais de

trinta anos. Esse autor explica que no início do século XIX existiam cerca de 100 periódicos, em 1830 este número aumentou para 500 e em 1850 registravam-se 1000 títulos. No ano de 1900 o número atingiu 10.000 títulos, e, segundo uma avaliação feita pela *Library of Congress* dos Estados Unidos, por volta da década de 1960 foram publicados mundialmente cerca de 30.000 títulos de periódicos técnicos e científicos.

Ainda, segundo levantamento de Targino e Garcia (2000), somente na *Science Citation Index* (SCI), da Base de Dados do *Institute for Scientific Information* (ISI), estavam cadastradas 16.000 publicações entre periódicos, livros e anais de congressos, além de 8.000 periódicos técnico-científicos.

No Brasil, segundo dados do CNPq[10], somente no período compreendido entre 2000 a 2003 foram publicados aproximadamente 894 mil trabalhos (artigos, livros, teses, dissertações e trabalhos em eventos) pela comunidade científica brasileira. Tais números são relevantes para esta pesquisa por dois motivos: por demonstrarem um grande volume de informação produzido e para reflexão acerca da forma pela qual é feita a comunicação e organização desta informação.

Como o volume de informação apresenta crescimento exponencial, é previsível que a produção científica nacional tenha crescido consideravelmente, aumentando ainda mais os estoques já acumulados ao longo do tempo. Sabe-se da necessidade de comunicação da ICT para que a mesma cumpra sua função social de compartilhar o conhecimento. Mas, será que a relação entre o aumento da produção de informação é proporcional à capacidade de comunicação? Em parte sim, e a comunicação científica cresceu também nos últimos anos, beneficiada pelo avanço das redes eletrônicas de comunicação.

Os dados apresentados pelo CNPq demonstram uma grande produção nacional de ICT, porém, ICT sem fluxo é estoque sem utilidade. Imagina-se que o ideal seria a existência da seguinte relação: quanto mais informação produzida, mais informação utilizada. Se essa proporção dependesse exclusivamente dos canais de acesso à informação, não haveria motivo para preocupação, pois os recursos eletrônicos tendem, cada vez mais, a ser massificados e barateados.

---

[10] Séries históricas dos Diretórios de Grupos de Pesquisa no Brasil: http://dgp.cnpq.br/censo2004/series_historicas/index_producao_cta.htm

Contundo, há questões a se ponderar. Uma delas é que os atores envolvidos nos processos de produção e uso da informação estão limitados à capacidade humana de leitura, interpretação e assimilação dos conteúdos. Machado (2003, p.71), após apresentar dados referentes ao crescimento mundial da produção de conhecimento, conclui que “o cérebro do homem não suporta o peso desse conhecimento acumulado e registrado em diferentes suportes”.

Além disso, para o contexto atual da ICT, o desafio maior não é somente oferecer mais alternativas de acesso à informação, pois tal solução envolve aspectos de telecomunicação e informática aplicados aos meios de comunicação. A preocupação maior diz respeito ao desenvolvimento de recursos que viabilizem, dentro de uma perspectiva de contínuo aumento dos estoques de ICT, aperfeiçoar os mecanismos que favoreçam a organização dessa informação, possibilitando que os estoques sirvam às finalidades para as quais foram concebidos.

Outro importante debate está relacionado às idéias defendidas por Ziman (1979, p.135) sobre um conhecimento científico público no qual “diferentes fragmentos de informação contidos nos diferentes trabalhos primários precisam ser reunidos e fundidos numa só peça, compondo uma coerente máquina”. O autor defende uma forma de comunicação de ICT capaz de estabelecer associações entre conhecimentos comuns em documentos diferentes.

Em trabalho recente, Marcondes, Mendonça e Malheiros (2005) discutem a comunicação da ICT ressaltando os ideais de Ziman. Esses autores sugerem que, através de novas estruturas para publicações eletrônicas seja possível potencializar as relações entre trabalhos publicados. Essas novas estruturas seriam baseadas em linguagens de marcação e em ontologias de domínios específicos.

Na prática, seria preciso criar um mecanismo automático que “compreendesse” o significado contido em partes de um documento e relacionar essas partes com as de outros documentos. Essa “compreensão” - baseada na estrutura do documento - é feita a partir da descrição formalizada de conceitos e relações de um delimitado domínio de conhecimento. Esse

relativo controle permite que os chamados “agentes inteligentes”[11] possam entender fragmentos de diferentes textos que abordem uma mesma temática.

Na comunicação da ICT, tão importante quanto a estrutura dos documentos, é o seu conteúdo. Desta forma, controlar a estrutura não assegura que haverá uma associação semântica entre os conhecimentos registrados. Para ilustrar a questão é exposto um exemplo: há um artigo científico sobre política de ICT na França, que está estruturado segundo um padrão específico que foi adotado por outro artigo que discutiu a política de ICT no Brasil. Os agentes inteligentes seriam capazes de “entender” que na metodologia de ambos os artigos adotaram-se a entrevista como recurso metodológico. Também seria possível identificar que a revisão teórica continha autores comuns.

De fato, a introdução dos agentes inteligentes significará um importante recurso, porém o conceito de semântica, no que se refere à estrutura do documento, difere de um outro conceito de semântica (o adotado por esta pesquisa) condicionado a significados. A atribuição de significados está delimitada por universos lingüísticos distintos, que definem as relações entre coisas e idéias sobre estas coisas. Este assunto será retomado nas seções 3 e 4. Por enquanto será discutido o desenvolvimento das políticas nacionais de ICT.

2.2 DESENVOLVIMENTO DAS POLÍTICAS NACIONAIS DE INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA E DOS SISTEMAS DE INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA

Seguindo ao debate sobre a comunicação da ICT, será contextualizado o desenvolvimento histórico das políticas de ICT brasileiras. No Brasil, o ambiente científico e tecnológico e, conseqüentemente as políticas relacionadas à C&T, caracterizam-se pela presença governamental. A esse respeito Bertero (1994, p.1) afirma que

> a gestão e a condução dos esforços que em nosso país objetivaram o desenvolvimento científico e tecnológico sempre foram de iniciativa governamental e consequentemente ocorreram num contexto de administração pública e com a presença quase exclusiva do Estado.

[11] Os Agentes Inteligentes são programas que coletam conteúdos informacionais em servidores da Web, processam e compartilhem os resultados com outros programas.

A opinião de Bertero é corroborada por outros autores, e há opiniões mais específicas relacionadas à ICT que demonstram a forte intervenção estatal na gestão da ICT no Brasil. Para Martins (2004) essa interferência de natureza pública (sobretudo na ICT) é recente e coincide com o desenvolvimento da área da Ciência da Informação no Brasil. Assim, ambas – as políticas de ICT e a Ciência da Informação - se desenvolvem sob o amparo de planos e programas que, a partir da década de 1960, passaram a ter forte influência norte-americana.

O retrospecto feito por Dias M.M.K. (2001), que abarcou as décadas de 1950 a 1990, serviu como base para explicar o desenvolvimento histórico da política de ICT no Brasil. Entretanto, as análises aqui feitas não se limitaram ao trabalho de Dias, pois a contribuição de outros autores foi fundamental para a elaboração desta seção. Inicia-se definindo, de maneira despretensiosa, dois termos bastante usuais nessa seção: **plano** e **programa**. Um plano é uma sistematização formal produzida a partir das discussões e atividades desenvolvidas num processo de planejamento, enquanto que o programa é um conjunto de projetos afins com relação a um objetivo maior.

Nas décadas de 1950 a 1960 deu-se início ao processo de institucionalização e intervenção direta do Estado na formulação de uma política de C&T e de ICT no país. Criou-se o CNPq, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), o Instituto Nacional de Tecnologia, o Instituto Brasileiro de Biblioteconomia e Documentação (IBBD) e os sistemas nacionais de informação especializada. A produção de ICT centrava-se nas universidades, institutos de pesquisa e empresas públicas de Pesquisa e Desenvolvimento (P&D). Nesta mesma década assistiu-se aos primeiros passos dos sistemas internacionais de informação.

Essa primeira fase, de aproximadamente 20 anos, teve momentos de avanços e retrocessos. Para Marques (1994), o governo de Getúlio Vargas demonstrou-se bem assessorado ao adotar medidas visando à criação de condições internas para o desenvolvimento endógeno da tecnologia. Os anos subseqüentes - nos governos de Café Filho, Juscelino Kubitschek e Jânio Quadros – foram avaliados por Valentim (2002) como estagnados no que respeita às políticas de C&T, principalmente devido à redução de investimento financeiro no setor. A situação se agravou com o êxodo de cientistas brasileiros

no mandato de João Goulart, e ainda mais, com o desrespeito pelo trabalho científico e muitas perseguições políticas no mandato de Castelo Branco.

Esse quadro revela um problema ainda vigente nas políticas públicas brasileiras: a descontinuidade de programas, ações, planos e até mesmo instituições. Para um produto como a ICT e seus respectivos sistemas, o valor cumulativo é fundamental, pois sua existência fundamenta-se em um processo de construção, em que as partes vão se encaixando até compor um todo no presente, que não se esgota, pois servirá como insumo para produzir a ICT futura. Assim, infere-se que o contexto em que a ICT brasileira nasce a torna fragilizada.

Na década de 1970 implantou-se uma política de C&T orientada pelos Planos Nacionais de Desenvolvimento (PND). No I PND, além do Banco de Patentes do Instituto Nacional de Propriedade Industrial (INPI), foi criado o Sistema Nacional de Informação Científica e Tecnológica (SNICT), passando o setor de ICT no Brasil a dividir-se em dois vetores: os sistemas surgidos após as iniciativas e esforços de integração e coordenação desenvolvidas pelo IBBD e os sistemas pertencentes a áreas estratégicas privilegiadas no planejamento econômico nacional, que eram coordenadas por uma política geral de ICT.

No II PND, o IBBD tornou-se o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e passou a planejar e coordenar o setor de ICT no país, assumindo várias missões do extinto SNICT. O IBICT inaugurou o primeiro curso de pós-graduação (mestrado) em Ciência da Informação no país.

Para Valentim (2002, p. 93), na década de 1970, marcada pelos governos de Médici e Geisel, houve

> [...] uma atenção especial para com o desenvolvimento científico e tecnológico, formulando uma política de C&T exposta em dois documentos: o I e II Plano Nacional de Desenvolvimento (PND) e o I e II Plano Básico de Desenvolvimento Científico e Tecnológico (PBDCT).

Apesar do fortalecimento do ambiente científico nacional nesse período, as críticas de Fonseca (1973, p.26) relativas à ICT são ríspidas ao dizer que

> [...] ninguém, no Brasil, está levando esse problema [da informação científica] a sério. Ninguém: nenhuma universidade, nenhuma academia, nenhum instituto. Os chamados “Serviços de Documentação” dos nossos ministérios e de outros órgãos – inclusive o do D.A.S.P – são uma farsa, bem típica da época em que foram todos criados: a do chamado *Estado Novo*.

Vê-se que Fonseca faz referências à documentação. Salienta-se que o termo documentação esteve durante muito tempo atrelado à ICT. Essa relação foi tão forte que Gomes (2006[12]) chegou a afirmar, no início da década de 1980, que a "informação científica, ou Informação Científica e Tecnológica – ICT, são expressões utilizadas como sinônimo de "documentação científica".

Neste mesmo texto, Gomes descreveu claramente sua preocupação com a adoção – no final da década de 1970 - de recursos informáticos nos serviços de informação brasileiros, e considerou essencial uma análise crítica destes recursos para evitar que novos erros fossem cometidos. Suas críticas recaíram, principalmente, na "importação" de tecnologias sem um estudo detalhado para saber se elas – as tecnologias - seriam adequadas ao contexto da ICT brasileiro.

Outros problemas críticos foram apontados por Gomes. Abaixo, são citados os mais condizentes aos interesses desta pesquisa:

- Os analistas brasileiros foram os principais agentes na venda dos "pacotes" de equipamentos e sistemas. Eles foram capacitados por fabricantes de computadores que mantinham, instalados no Brasil, equipamentos para outras finalidades não relacionadas à ICT;

- a manutenção dos acervos (que sofria com a insuficiência de verbas para seu crescimento) foi prejudicada com a compra dos computadores, com manutenção dos equipamentos, e também com os altos salários dos analistas que passaram a comprometer consideravelmente os recursos financeiros;

- houve sub-utilização dos computadores, que foram inicialmente utilizados somente para cadastrar registros sem o propósito de servir à recuperação da informação;

- a automação dos serviços não foi procedida de um estudo que identificasse a real necessidade de sua adoção;

- os primeiros esforços se concentraram no uso de pacotes que adotavam a linguagem natural. Por desconhecer conceitos de organização da informação, os analistas desconsideraram recursos bibliográficos destinados ao tratamento da informação. Assim, foram geradas inconsistências nos serviços disponíveis para as coleções multidisciplinares, cuja linguagem é

[12] Em texto publicado originalmente em: GOMES, H. E. Informação científica. **Palavra-chave**, São Paulo, n.1, p.19-20, 1982.

fortemente caracterizada pela ambigüidade dos termos. Gomes frisou, já em 1982, que esses serviços apresentavam sérios problemas de recuperação e ressaltou que "os sistemas internacionais utilizam vocabulário controlado e não linguagem natural para recuperação, pois aquele é o instrumento indispensável para permitir a participação de países de diversas línguas" (GOMES, 2006).

De 1982 (ano em que Gomes publicou suas opiniões) aos dias atuais, as dificuldades para manter os SICTs organizados ainda são marcantes. No caso específico das análises feitas na **PL** identificaram-se problemas motivados pela ausência de controle na alimentação do sistema. Mas tal questão será posteriormente mais bem explorada.

Foi nos anos da década de 1980 que se instituíram, através do III PND, o Sistema Nacional de ICT e os centros especializados. Criaram-se ainda os Sistemas Estaduais de Informação Científica e Tecnológica (SEICT) e bancos de dados nacionais.

Nesse período, orientado pelo III PND, foi implantado o III Plano Básico de Desenvolvimento Científico e Tecnológico (PBDCT), 1980-1985. Para operacionalizar o III PBDCT, o CNPq planejou ações para quase todas as áreas do conhecimento científico e tecnológico. Entre elas, constou a *Ação Programada de Comunicações, Eletrônica e Informática*, com linhas específicas para as bases de dados.

Almejava-se o crescimento do mercado de bases de dados a partir do estímulo ao uso dos serviços de consulta a outras bases de dados e também a instrumentalização e operacionalização, no país, de bases de dados estrangeiras bem como o fortalecimento e aprimoramento das iniciativas das bases de dados nacionais (Citado[13] por Amaral, 1995, p.226).

Outro importante programa dessa década foi o Programa de Apoio ao Desenvolvimento Científico e Tecnológico (PADCT), iniciado em 1984 com o objetivo de "ampliar, melhorar e consolidar a competência técnico-científica nacional no âmbito de universidades, centros de pesquisa e empresas, mediante financiamento de projetos integrados" (IBICT, 1993, p.47).

Na visão de Valentim (2002) o PADCT poderia financiar projetos para a criação de bases de dados ou portais de informação, já que uma das

[13] BRASIL. Presidência. Secretaria de Planejamento. Plano básico de desenvolvimento Científico e Tecnológico, III. 1980-1985: comunicações, eletrônica e informática. Brasília, 1984. 186p. (Ação Programada em Ciência e Tecnologia, 17).

atribuições do PADCT era aperfeiçoar a infra-estrutura de apoio e serviços à C&T nacional. Infelizmente, “apesar de o programa ser estruturado em vários subprogramas, nenhum deles foi especificamente direcionado à consolidação de dados ou informações produzidas no país” (VALENTIM, 2002, p.95).

A década de 1980 também ficou assinalada por crises e instabilidade na ICT em razão de o Ministério da Ciência e Tecnologia (MCT) absorver muitas das funções do CNPq. O reflexo se deu principalmente porque o IBICT, até então um órgão vinculado ao CNPq, era responsável por muitas das ações da ICT nacional. A partir dessas alterações políticas iniciou-se para o IBICT

> um período de difícil transição no qual se destaca a rotatividade de seus dirigentes. Tal fato pode ter ocasionado uma possível descontinuidade administrativa, [...] resultando em interrupção total ou parcial de projetos, ou na geração de novas ações sem uma adequada análise dos produtos/serviços e dos impactos resultantes, principalmente junto aos usuários potenciais” (CUNHA, 2005, p. 7).

Concordamos com as opiniões de Cunha a respeito das possíveis descontinuidades provocadas por esse tumultuado período. E, é provável que, na década de 1980, as articulações necessárias para o avanço da ICT nacional tenham se fragilizado com a conjuntura vivenciada pelo IBICT. Talvez, se problemas de descontinuidade política fossem reduzidos, possivelmente os SICTs atuais refletissem ações bem sucedidas do passado.

A situação atual dos SICTs em nações desenvolvidas é um reflexo de uma infra-estrutura iniciada há décadas. Silva (1997) demonstra que a política da União Européia, para alcançar um domínio da ICT, já previa em 1973 um plano de ações que, em 1980, permitiria a implantação da Rede Euronet/Diane, composta por 60 centros distribuidores espalhados em 12 países, cobrindo cerca de 300 bases de dados. Planejamentos dessa natureza contribuem não apenas para um bom funcionamento sob o ponto de vista técnico, mas também para que esses sistemas sejam incorporados como elementos fundamentais na formulação de ações no âmbito da C&T.

A década de 1990 assinalou a mudança de paradigma em razão do rápido avanço das tecnologias da informação provocada pela popularização da Internet. Essa mudança teve origem em 1989, quando Tim Berners-Lee propôs a criação da WWW[14]. No Brasil, no âmbito da ICT, foram instalados os Núcleos

[14] Segmento da Internet composto de textos, sons e imagens que conjuntamente facilitaram a interação comunicativa entre os usuários. A WWW ficou tão conhecida que comumente é entendida como um sinônimo da Internet.

de Informação Tecnológica (NIT) e criados sistemas e/ou redes responsáveis pelo programa de disseminação da Informação Tecnológica.

Segundo Fujino (2004), na década de 1990 o governo buscou estimular a inovação e a parceria entre universidades e empresas, e um dos mecanismos que auxiliaria na criação de elos no contexto da C&T seria a ICT. Porém, uma pesquisa realizada em 2000 no serviço “Disque Tecnologia” da USP demonstrou que atividades fundamentais, como a inclusão de mecanismos de difusão e transferência da informação, não foram contempladas em instituições de ensino e de pesquisa.

Nota-se que houve uma expectativa equivocada de estímulo ao uso dos recursos de ICT a partir da criação de NITs. Prever que a ICT será invariavelmente bem utilizada pelos atores da C&T, em razão de necessitarem dela como insumo à produção do conhecimento, é um raciocínio falacioso. É imprescindível considerar que outras variáveis, além das questões técnicas, interferem na resolução de problemas de informação tecnológica.

Foge ao foco desta pesquisa aprofundar a questão, mas é oportuno ressaltar que as políticas econômicas brasileiras, há décadas, privilegiam o lucro nos mercados financeiros e não estimulam os investimentos em P&D. Assim, se os lucros podem ser alcançados a curto prazo com os juros, inflação e mercado de ações, para quê desenvolver avanços tecnológicos para fins produtivos?

Em estudo da CNI/SENAI (1996), verificou-se nas empresas brasileiras (micros, pequenas, médias, e grandes) uma baixa utilização de normas técnicas (micro 10,5%; pequena 14,3%; média 31,7%; e grande 43,9%) e de Bancos de Dados de patentes/propriedade industrial (micro 7,7%; pequena 4,4%; média 7,7%; e grande 16,8%). Basicamente, o uso de informações era limitado àquelas disponíveis na própria empresa (micro 59,9%; pequena 63,0%; média 65,1%; e grande 72,8%). Esses números evidenciam o quanto as deficiências nacionais relacionadas à ICT não representam um problema que possa ser resolvido somente a partir de soluções técnicas.

Desta forma, registra-se um questionamento: os produtores e usuários da ICT no Brasil estão suficientemente familiarizados com os SICTs a ponto de usá-los de forma mais adequada? Ressalta-se que o uso engloba tanto a

busca por informações em SICT como também a livre inserção de informações em alguns sistemas, como é o caso da **PL**.

A propagação de sistemas eletrônicos de ICT disponíveis online na década de 1990 foi bem acentuada. Destaca-se, deste período, o projeto para o desenvolvimento de uma metodologia para armazenamento, disseminação e avaliação de publicações científicas em meio eletrônico: a Biblioteca Científica Eletrônica Online (SciELO[15]). O projeto, resultado de uma parceria entre a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), o Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME) e editores de periódicos científicos, desenvolveu uma metodologia própria que buscou atender à

> demanda de editores científicos por soluções confiáveis para a publicação eletrônica de seus periódicos compatíveis com as iniciativas internacionais mais importantes; [e] uma antiga demanda referente à operação de bases de dados bibliográficos para não apenas controlar e disseminar a literatura científica, mas também permitir a produção de indicadores para subsidiar estudos de bibliometria, informetria e cienciometria sobre a produção científica nacional relevante" (PACKER, 1998, p.114).

A metodologia da SciELO estabelece um controle para a disponibilização das informações em seu sistema. Além disso, todos os documentos que compõem o seu acervo passam pelo crivo de conselhos editoriais dos periódicos disponibilizados. Esse tratamento dado à informação permitiu ao SciELO oferecer índices (autor, título, resumo, assunto, afiliação – país/instituição, ano de publicação e tipo de artigo) que mantêm um razoável controle nos seus termos. A partir desse controle viabilizam-se buscas por termos que representam os documentos, conforme análise de F.M. Silva (2002), que interpretou a produção científica do Periódico Ciência da Informação a partir do índice de assuntos disponível na SciELO.

Para os interesses desta pesquisa, destaca-se que nos anos 1990 o movimento de defesa ao livre acesso à ICT ganhou força através de redes eletrônicas de comunicação. A esse respeito, Sena (2000) aponta que os Arquivos Abertos seriam uma alternativa para a comunicação científica brasileira. No mesmo sentido, Triska e Café (2001, p,92), ao descreverem os Arquivos Abertos como sub-projeto da Biblioteca Digital Brasileira (coordenado

[15] http://www.scielo.br

pelo IBICT), afirmam que as motivações para a implantação desses sistemas foram:

- ampliar a visibilidade nacional e internacional da produção intelectual brasileira em C&T;
- melhorar o fluxo de comunicação científica e tecnológica;
- incrementar o ciclo de geração de novos conhecimentos.

Concordamos totalmente com a idéia de se sociabilizar integralmente o conhecimento científico, mas chamamos novamente a atenção para o fato de que disponibilizar e acessar documentos eletrônicos depende de soluções tecnológicas, enquanto que o fluxo e a comunicação da informação exigem outras ações que envolvem processos de tratamento e organização da informação.

A opinião de Marcondes e Sayão (2001, p.25) é bastante coerente quando diz que

> somente a disponibilidade de textos brasileiros em C&T online não teria grande impacto sobre a comunicação científica e a ciência no país sem a existência de serviços de informação que viabilizem o acesso de forma fácil a estes conteúdos.

A avaliação dos SICTs brasileiros exigiu uma compreensão sistêmica dos fatos ocorridos até o presente momento. Por compreensão sistêmica, entenda-se uma análise que não se limita a um sistema computacional, a uma iniciativa institucional, ou a um programa governamental. A visão sistêmica contempla um conjunto inter-relacionado de fatores sociais, políticos, técnicos e econômicos.

A literatura demonstrou que para observar esse conjunto de fatores deve-se evitar uma leitura ahistórica dos fatos, pois o estado atual reflete uma história da ICT brasileira marcada por ações equivocadas que desfavoreceram a consolidação de um sistema sólido e duradouro.

Foi desfavorável, por exemplo, o Brasil ter iniciado um processo de institucionalização da ICT somente em meados das décadas de 1950 a 1960 através da criação do CNPq e do IBBD. Contudo, mais prejudicial foi a descontinuidade dos programas e projetos (em geral de natureza pública) ao longo dos anos. A descontinuidade de ações demonstrou que a C&T (e consequentemente a ICT) jamais alcançou o status de assunto prioritário nas mesas de discussões das esferas governamentais.

Infelizmente, prevaleceram nos meios produtivos brasileiros: a indiferença ao avanço tecnológico; a valorização da mão-de-obra barata; e o

interesse por ganhos no mercado financeiro em períodos de altos índices inflacionários. Assim, a perspectiva de uma rentabilidade a curto prazo não resultou benéfica para a P&D brasileira, e pouco importava a existência de SICTs bem estruturados neste contexto, porque a produção e uso da ICT se tornaram secundárias.

Almejar o êxito de toda a ICT num país de dimensões como o Brasil é uma meta ambiciosa. Mesmo em países que alcançaram avanços na produção e organização da ICT, é improvável imaginar um conjunto coeso e unificado de atores e instituições cooperando mutuamente. Entretanto é possível estabelecer princípios mínimos que tornem mais estáveis os serviços e produtos informacionais, e para tal, é necessário formular políticas com base em cenários que antecipem o crescimento das demandas de uso e produção dos estoques.

Prever o crescimento e possibilitar a interação entre os sistemas requer níveis de controle para que haja elementos comuns entre os sistemas. Esse nível de controle variará conforme o contexto de uso e produção da informação. Por isso, será sempre importante estar atento aos padrões de organização e classificação de informações ou padrões técnicos de comunicação, também conhecidos como protocolos.

Nos EUA e na Europa a preocupação em estabelecer e manter princípios mínimos de organização da informação e políticas para a ICT já motivava discussões desde o final do século XIX. Assim, a adoção das tecnologias eletrônicas no gerenciamento de sistemas de informação não ocorreu casualmente nesses países.

No Brasil, as primeiras experiências de uso dos recursos eletrônicos na ICT serviram para expor a fragilidade da organização da informação, demonstrando o quão mal planejados e desorganizados estavam. Gomes (2006) enfatizou que o Brasil foi mero comprador de pacotes prontos (e caros) de programas. E chamou a atenção para o fato de que enquanto sistemas internacionais adotavam vocabulários controlados uniformes para fins de recuperação da informação, no Brasil usava-se, indiscriminadamente, a linguagem natural na automação dos sistemas.

Foram observados, no final da década de 1990, problemas com a popularização da Internet, pois houve uma convergência não planejada para a

criação de produtos e serviços de ICT acessíveis via WWW. Desta vez, o Brasil já havia alcançado autonomia no desenvolvimento de TIC, e não dependeu de apoio externo para criar seus próprios sistemas computacionais. Porém, negligenciou-se a necessidade de políticas de organização da informação, e assim esforços isolados (e provavelmente desnecessários) produziram estoques com padrões próprios e com recursos não compartilháveis de informação.

O planejamento e implementação de recursos e serviços da ICT exigem uma prévia identificação das necessidades e competências informacionais dos atores da ICT nacional que utilizarão tais recursos. Desenvolver tais sistemas, pressupondo-se serem úteis e de uso natural por parte da comunidade científica, é um modelo que ganhou espaço como uma nova forma de lidar com a informação. E os sistemas abertos, que transferem para os usuários a responsabilidade de alimentar os estoques, são um sinal de que a comunidade de C&T começa a incorporar essa nova modalidade para disseminação e uso de suas produções.

Neste modelo aberto os estoques são constituídos com um custo relativamente baixo. Assim, acervos crescem com investimentos menores, comparando-se a modelos caracterizados pela presença de intermediários entre produtores de conhecimento e estoques. Essa relativa vantagem econômica pode comprometer a finalidade do sistema, como por exemplo, a da **RI**. Porém, o benefício trazido pela racionalidade econômica no crescimento dos estoques – ainda que traga perdas na organização da informação - talvez represente o início de um novo capítulo da história da ICT nacional.

3 A ORGANIZAÇÃO DA INFORMAÇÃO

Nesta seção são abordados os fundamentos gerais da organização da informação, particularmente a organização da informação em meio eletrônico. Em seguida, será discutida a temática da organização da informação em contextos digitais, que é o objetivo da pesquisa. Nesse aspecto interessa discutir os recursos eletrônicos voltados à representação (temática e descritiva) mais importantes para a organização da ICT.

3.1 DELIMITAÇÃO DE CONCEITOS

Debater a organização da informação exige a delimitação de conceitos relacionados à temática investigada nesta pesquisa. Inicialmente, expõem-se concepções a respeito de dois termos: a informação e o conhecimento.

Para Le Coadic (2004, p.4), a informação é

> [...] um conhecimento inscrito (registrado) em forma escrita (impressa ou digital), oral ou audiovisual, em um suporte. A informação comporta um elemento de sentido. É um significado transmitido a um ser consciente por meio de uma mensagem inscrita em um suporte espacial-temporal [...].

Percebe-se que uma das características da informação é que a mesma deve estar explicitada, diferentemente do conhecimento que é individual, subjetivo e produzido a partir da assimilação da informação. A este respeito, Svenonius (2001) apresenta uma visão dicotômica da informação: uma visão fundamentada na Teoria da Informação, referente a aspectos mensuráveis da informação contida em mensagens e a outra, voltada para o conteúdo que a mensagem carrega. Nessa segunda visão, a informação pode ser organizada, desde que registrada.

Entende-se que a informação tem condições de produzir conhecimento, mas esse "só se realiza se a informação é percebida e aceita como tal e coloca o indivíduo em um estágio melhor de convivência consigo mesmo e dentro do mundo em que sua história individual se desenrola" (Barreto, 1994, p.3).

Além do mais, em outro artigo, Barreto (1999) destaca que a produção da informação é operacionalizada através de técnicas que envolvem "atividades de reunião, seleção, codificação, redução, classificação e armazenamento de informação". Ainda segundo o autor, o processo de produção da informação é capaz de criar estoques, que são potenciais

geradores de novos conhecimentos, mas é necessário que eles não fiquem estáticos, isto é, que sejam utilizados, caso contrário serão meros repositórios de documentos. A visão de Barreto vem ao encontro das preocupações sobre o crescente interesse em criar repositórios de informação (sobretudo eletrônicos) sem definir com precisão como esses estoques se inscreverão em fluxos.

Delimitadas as concepções sobre a informação e o conhecimento, abordar-se-á a organização da informação; para tanto, inverteremos o processo apresentando, inicialmente, conceitos que não competem às discussões tratadas neste estudo.

Apesar de usualmente adotado pela área da Ciência da Informação, o termo organização da informação é também utilizado por outras disciplinas, porém com sentidos diferentes, relacionados a objetos que não condizem com as discussões desta tese. Na área da Arquitetura, o termo expressa a forma usada por um programa computacional (o CAD) para apresentar imagens aos usuários, ou seja, como as imagens (que são consideradas informações) podem ser distribuídas de forma mais organizada na tela do computador (GIACAGLIA, 2001).

Em outro estudo, desta vez da área da Educação, Dias (2000) explora o uso do hipertexto no ensino e aprendizagem, e denomina organização da informação a estrutura na qual estão dispostos os conteúdos de um documento, envolvendo basicamente aspectos estéticos e de relações (links) entre partes do documento. Discussão semelhante foi observada em Mander, Salomon e Wong (1992) na área da Computação.

Tais exemplos demonstram que o entendimento sobre a organização da informação precisa ser definido, pois uma definição objetiva evitará inconsistências conceituais comuns, quando se mesclam abordagens de áreas distintas a respeito de um objeto.

Outro aspecto importante é que não será adotado, nesta pesquisa, o termo *organização do conhecimento*, para evitar discordâncias terminológicas, ainda que se tenha percebido autores que o utilizam (referindo-se a conceitos análogos à organização da informação ou a aspectos filosóficos e epistemológicos do conhecimento).

Então, o que é a organização da informação? Apesar de ser um termo recorrente, não é fácil identificar na literatura científica delimitações claras

sobre esse objeto. Em alguns autores é possível perceber, implicitamente, a concepção que eles têm da organização da informação, mas uma definição feita sistematicamente por enunciados não é comum.

Alvarenga (2003, p.12) afirma que a organização da informação compreende um processo de representação destinado

> prioritariamente à recuperação eficaz por parte dos usuários. Para que tal ocorra torna-se necessário que profissionais da informação desenvolvam e implementem sistemas representacionais que estabeleçam a confluência entre a organização cognitiva imposta ao conhecimento pelo seu produtor (representação primária) e a organização conceitual imposta ao documento pelo especialista da informação (representação secundária).

Nesse mesmo artigo, a autora menciona outras vezes a organização da informação, como no trecho referente ao trabalho publicado por Shera, na década de 1950, que tratava do aperfeiçoamento da organização da informação gráfica e explorava o processo de formação de conceitos no cérebro humano. O texto de Alvarenga (2003) ainda que preciso, contribuiria ainda mais para a discussão do tema se incluísse uma definição objetiva de organização da informação.

Já o artigo de Tristão et al (2004) traz em parte do título o termo *organização da informação*, porém, o mesmo não é mencionado explicitamente em nenhum outro momento do documento. E adota-se o termo organização do conhecimento em todo o texto, sem justificar a razão dessa mudança. Esse fato não compromete a qualidade do trabalho e, assim como no texto de Alvarenga (2003), o leitor consegue compreender que os autores discutem aspectos da organização da informação para fins de recuperação.

Para Tristão et al (2004, p.3) o conhecimento registrado e publicado, ou seja, a informação em nossa concepção, tem ao seu dispor sistemas de organização que

> [...] existem desde os tempos remotos e estão presentes em todas as áreas do conhecimento humano, de modo simples aos mais complexos. Esses sistemas abrangem: classificações, tesauro, ontologias, glossários, dicionários, enciclopédias, guias, específicos a cada área e, em sua maioria, ligados às bibliotecas e outras organizações de gerenciamento da informação visando organizar, recuperar e disseminar a informação.

Não é rara a carência de definições em temáticas pesquisadas pela área da Ciência da Informação. Além das dificuldades para delimitar o tema organização da informação, percebe-se, através do levantamento de Smit, Kobashi e Tálamo (2004) que a organização não é “conceituada como um

meio, mas como um fim em si". Essas autoras explicam que não somente a organização, mas também o tratamento da informação é visto como um conjunto de procedimentos. Assim, diante da dificuldade em encontrar definições sobre a organização da informação, optou-se por estabelecer uma definição própria voltada aos propósitos desta pesquisa.

Como já se definiu a informação no início desta seção, é desnecessário fazê-lo novamente, mas enfatiza-se que ela é registrada, contida num suporte, num documento que contém o conhecimento de um ou mais indivíduos e já está estruturado em modelos socialmente aceitáveis de comunicação, codificados numa linguagem compreensível por aqueles que a produzem e a assimilam.

Segundo Svenonius (2001) organizar a informação tem como objetivo essencial agrupar informações por semelhança, ou visto de outra forma, separar as informações que são diferentes. A organização da informação está relacionada ao documento, busca ordená-lo, arranjá-lo para torná-lo disponível. E o esforço para estabelecer esta disposição principia num propósito, o de estabelecer meios para encontrar a informação. Estes meios (ou procedimentos) são reconhecidos no processo de tratamento da informação, tratamento este que não pode alterar o documento, porém criar novas informações a partir dele. Desta maneira, o processo de tratamento de informação é feito a partir de metodologias formalizadas de reconhecimentos temáticos e descritivos do conhecimento inscrito num suporte documental.

A organização da informação não se limita a um conjunto de procedimentos, mas se realiza através deles a partir do tratamento da informação, que é orientado por recursos[16] bem definidos, adotados por instituições ou ambientes específicos destas instituições. A relação entre a organização da informação e o tratamento da informação não é meramente técnica, pois o ato de organizar não é um meio que se justifica pelo fim, porém requer meios que possibilitem a esse ato alcançar seu objetivo maior, que são os fluxos de informação. Desta forma, a organização da informação tem como finalidade o fluxo ou uso dos estoques de conhecimento (registrados) e, para

[16] Processos, métodos, e produtos como a indexação, catalogação, classificação, tesauro, vocabulários, índices, listas, dicionários, etc

tanto, deve tratá-los a partir dos domínios de conhecimento nos quais eles foram gerados e serão utilizados.

3.2 A INFORMAÇÃO EM MEIO ELETRÔNICO

O conhecimento armazenado em meio eletrônico diz respeito a um conceito físico do fluxo de elétrons, o que implica uma diferença entre o meio (eletrônico) e seu conteúdo. Para uma melhor explicação, imagina-se duas versões da Bíblia, uma impressa e a outra digital. Sabe-se que em ambas o texto trará igualmente ensinamentos da moral cristã. Sabe-se também que o conteúdo da versão impressa está em um suporte físico composto de átomos, enquanto que a digital está em um meio eletrônico composto de elétrons. Logo, analisar a organização da informação que ocorre em meio eletrônico não equivale a investigar o meio eletrônico da informação.

Outro detalhe a ser esclarecido é quanto aos termos usualmente relacionados ao meio eletrônico. O mais elementar é o **bit**, que é a menor unidade do meio eletrônico, legível por máquinas e de natureza binária, composto apenas por 2 elementos: 0 (zero) e 1 (um). Esse universo binário sustenta a existência dos ambientes digitais e é uma representação numérica da realidade acessível através de equipamentos eletrônicos. Desta forma, ao ler um texto armazenado eletronicamente, a versão apresentada na tela do computador resulta de um processamento lógico de bits registrados em um arquivo digital. A máquina converteu dados numéricos armazenados em signos compreensíveis humanamente, através das chamadas interfaces.

Apesar do caráter técnico, a explicação justifica a adoção da expressão '*Informação em meio eletrônico*' ao invés de '*informação eletrônica'*. A segunda alternativa é considerada mais ambígua, já que pode ser entendida tanto como a informação materializada em bits (o que não seria exatamente a informação e sim um arquivo eletrônico) ou qualquer atividade relacionada à informação a partir do uso de dispositivos eletrônicos. Por outro lado, a expressão 'Informação em meio eletrônico' parece mais adequada aos processos da informação utilizando-se os meios eletrônicos. Por considerá-la menos dúbia, ela será utilizada.

Outra importante delimitação terminológica refere-se aos recursos eletrônicos relacionados à informação, mais precisamente aos instrumentos tecnológicos utilizados no contexto da informação em meio eletrônico. Genericamente, são denominados pela expressão Tecnologia de Informação e Comunicação (TIC), comum para o conjunto que engloba equipamentos, processamento e transmissão de dados eletronicamente.

A informação em meio eletrônico vem despertando o interesse no desenvolvimento de novas metodologias orientadas aos três elementos dos processos da informação (LE COADIC, 2004): a construção, a comunicação e o uso. Um aspecto a se destacar desse interesse recai no fato de que nem sempre a adoção de modelos usuais de organização da informação em mídia impressa serve aos propósitos do contexto digital. Um exemplo foi a avaliação negativa de Sondergaard et al (2003) com relação ao modelo de comunicação de ICT elaborado pelo UNISIST em 1971. O UNISIST foi um programa intergovernamental liderado pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) para cooperação no campo da ICT, que durou 4 anos para ser proposto. Segundo os autores, o citado modelo requer uma revisão e atualização para se adequar às transformações ocorridas nos processos de comunicação. E há duas razões de mudanças, uma teórica e a outra empírica.

Sob o ponto de vista teórico, a crítica recai na viabilidade de um modelo único para comunicação científica que seja adequado para lidar com uma produção tão ampla de conhecimento feito por áreas tão distintas. Para os autores, as Ciências Humanas e Sociais são as que menos se enquadrariam no modelo, consequentemente seriam as mais desfavorecidas. A razão empírica é mais simples de ser compreendida, pois se refere às mudanças ocorridas no processo de comunicação científica a partir do avanço das tecnologias de comunicação e informação nos 32 anos que separam a proposta do UNISIST do estudo de Sondergaard et al (2003).

Barreto (1998,1999) defende as modificações nas estruturas de informação e conhecimento, que passaram a exigir novas abordagens nas discussões sobre os fluxos informacionais em meio eletrônico. Para este autor, na interação do receptor com a informação,

> o receptor da informação deixa a sua posição de distanciamento alienante em relação ao fluxo de informação e passa a participar de sua fluidez como se estivesse posicionado em seu interior. Sua interação com a informação é direta, conversacional e sem intermediários (Barreto, 1998, p.125).

De fato, grandes transformações ocorreram no ambiente informacional, contudo discorda-se deste autor quando o mesmo critica negativamente o que ele chama de *rituais de ocultamento da informação*, referindo-se às formas usuais de tratamento da informação. Esses rituais nada mais são do que a adoção de instrumentos de metalinguagem e universos semânticos privados, segundo ele próprio. Barreto julgou precipitadamente que a adoção de recursos eletrônicos no âmbito da informação seria suficiente para estabelecer um paradigma totalmente diferente no contexto do tratamento da informação, mas sabe-se que isso não ocorreu.

Na visão de Barreto (1998, p.126), aqueles que defendem o uso desses recursos (os tradicionais), mantêm uma ideologia envelhecida, que representa um "*entrave ao desenvolvimento do pensamento e ao livre fluxo da informação*". As ontologias, os metadados, e a web semântica (que serão discutidos no item 3.4 a seguir) parecem desmentir tais considerações, pois não representam um paradigma revolucionário nas formas de tratar a informação já que mantêm relações com recursos menos atuais.

Uma outra opinião análoga, porém menos radical, pode ser observada em Robredo (2005, p.253), ao retratar que

> o crescimento exponencial da informação científica e técnica – e da informação em geral - e as crescentes facilidades de comunicação e difusão de todas essas informações, por meio dos novos canais e mídias que foram surgindo, tornaram inadequado e obsoleto o modelo até então vigente.

Sobre esse modelo, entende-se que o autor se refere aos processos mais tradicionais de produção, registro, armazenamento e tratamento da informação, já que, em seguida, Robredo discute temáticas como bibliotecas digitais, metadados, repositórios online, etc.

As mudanças no âmbito da informação em meio eletrônico também estão presentes em outros estudos. Um deles, não especificamente concernente à questão do eletrônico, é o trabalho de Rayward (1994) que, como é próprio deste autor, traz uma instigante contextualização histórica. Ele trata do suporte da informação, como meio de mobilização do conhecimento e

revela que há uma problemática que vem se desenvolvendo a partir do despreparo em lidar com outras mídias além da impressa.

Uma outra linha de discussão, apontada por Davenport e Cronin (1989), trata de um recurso bastante explorado pelas novas investigações sobre o fluxo informacional: o hipertexto. Não se avançará neste item, mas vale a pena mencioná-lo pela sua relevância e intensa aplicabilidade em muitos dos sistemas eletrônicos de informação.

O interesse despertado para estudos sobre o meio eletrônico proporcionou uma gama variada de caminhos para pesquisas, muitos deles equivocados. Autores como Shera, Landau, Cleveland e Foskett alertaram, desde os anos 1970, quanto à inconsistência dos estudos desenvolvidos pelas áreas da informação sobre os sistemas informatizados. Foskett percebia o risco dos rumos das pesquisas da área, por elas estarem "*reduzindo [a informação] a commodities, com ênfase na tecnologia do processamento da informação sem olhar para o seu significado ou destino*" (citado por PINHEIRO E LOUREIRO,1995, p.46). Para Foskett, a ênfase na técnica negligenciou o conteúdo, o foco nos sistemas e produtos gerou um fosso que deveria ter sido preenchido por disciplinas como a Ciência da Informação. Num exemplo elementar, é como se a medicina esquecesse de tratar do corpo para desenvolver sofisticados instrumentos de diagnóstico.

Outro problema já mencionado é a facilidade encontrada para alimentar as bases de ICT. Esta simplificação despertou o interesse pela produção e disponibilização de conteúdos nos canais eletrônicos de comunicação, gerando um volume heterogêneo e não sistematizado de dados. Desta forma, numa rede como a Internet, segundo Barreto (2000), é mais fácil encontrar informações que satisfaçam necessidades básicas do sujeito, pois uma modalidade de informação que resulta da reflexão, criatividade, realização profissional e pessoal, agregando maior valor qualitativo, é menos visível aos sistemas automatizados de busca.

De acordo com Alvarenga (2001) "o volume de informações livremente colocado na web torna impossível um tratamento da informação nos moldes tradicionais". A autora destaca que a Ciência da Computação vem empreendendo esforços no sentido de automatizar a classificação dos objetos em meio eletrônico, porém, faz uso da linguagem natural na representação e

recuperação da informação, o que dificulta simbolizar conceitos, restringindo-se apenas a unidades lexicais.

Um dos segmentos mais importantes da informação em meio eletrônico é o da Recuperação da Informação (RI) - inclusive uma das finalidades da **PL** é proporcionar a busca por currículos (e seus respectivos conteúdos) através de um sistema elaborado para este fim. Além disso, o próprio ato de organizar a informação está intrinsecamente atrelado também à RI, que será discutida a seguir.

3.2.1 A Recuperação da Informação

O termo Recuperação da Informação, criado por Calvin Mooers em 1951, refere-se aos “aspectos intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação” (Mooers citado por Ferneda, 2003, p. 11).

Essa definição é complementada por Kent (1972), ao dizer que recuperar informação, diferentemente da idéia de se recuperar qualquer outro objeto, não faz referência à procura de algo perdido, porém, é a busca por algo que foi (antecipadamente) organizado para ser encontrado.

A data de criação do termo RI é contemporânea ao lançamento do primeiro computador (o ENIAC) em 1946, logo, os princípios de RI independeram do processamento eletrônico de dados, pois os primeiros computadores limitavam-se à execução de operações matemáticas. Além do mais, é vasta a literatura em áreas como a Documentação e Biblioteconomia que discute aspectos de RI antes da primeira menção a este termo no ano de 1951.

Numa compreensão mais atual, Ingwersen (1992) entende que a RI contempla processos de representação, armazenamento, busca e encontro de informação relevante. Para o autor (baseado em Van Rijsbergen), a relevância da informação é a medida, ou grau de correspondência ou utilidade existente entre um texto (ou documento) e uma questão (ou requisição) de informação por uma pessoa.

No mesmo trabalho, Ingwersen (1992, p.58) explora três importantes abordagens da RI: a tradicional, a orientada ao objeto, e a orientada ao usuário. A seguir, apresenta-se uma síntese das três abordagens:

a) abordagem tradicional: é chamada de tradicional por considerar que suas origens remontam às primeiras técnicas utilizadas para organização da informação (como teorias de classificação e indexação adotadas pela Biblioteconomia). Volta-se ao refinamento das técnicas de RI e a métodos de representação, envolvendo testes científicos controlados e problemas de relevância. As soluções são, maiormente, *ad hoc*[17], e empregam técnicas de algoritmos para análise de texto. É centrada em questões atinentes à ICT e, consequentemente foca-se nos atores da C&T. Apóia-se em disciplinas como a Matemática, a Lingüística, a Ciência da Computação e a Inteligência Artificial.

b) abordagem orientada ao usuário: originou-se entre os anos de 1970 e 1980, focando aspectos psicológicos e comportamentais da comunicação entre usuários e produtores da informação. Destaca-se por buscar uma compreensão mais profunda das atividades executadas pelos intermediários. Fundamenta-se no ASK[18] (Estado Anômalo do Conhecimento), descrito por Belkin (INGWERSEN, 1992, p.28), que estuda o comportamento do usuário e suas necessidades informacionais, incluindo situações na vida real e condutas comportamentais. Propõe modelos e tipologias de usuários, estuda interações entre o usuário e intermediários (humanos ou automatizados) e também o desenvolvimento de interfaces básicas e modelos de busca por meio de entrevistas. Centra-se no que Ingwersen denominou de informação vital para a sociedade, que em outras palavras, abrange usuários de quaisquer níveis sociais. Apóia-se na Psicologia Cognitiva, Psicolingüística e Sociologia.

c) abordagem cognitiva: tem origem na década de 1980 e enfatiza atividades cognitivas (atividades mentais, emocionais, motivacionais) com um forte viés para análises individuais na interação entre os sujeitos e os sistemas de RI. Entende a RI como um processo que envolve, além dos estados

[17] Se novas buscas são submetidas a um SRI e o acervo de um SRI é pouco modificado, denomina-se a operação de "recuperação ad hoc". Por outro lado, se as buscas se mantêm relativamente estáticas enquanto novos registros são adicionados, chama-se essa operação de filtragem (filtering). A recuperação ad hoc é comum na maior parte das buscas em SRI, enquanto que a filtragem ocorre frequentemente em atividades de monitoramento de fontes informacionais. (Souza, 2006)
[18] Do inglês Anomalous State Knowledge.

cognitivos, interações complexas, modelagem de tarefas e domínios cognitivos (trata-se de uma RI baseada em conhecimento). Tal abordagem busca uma RI "inteligente", composta pela unificação de diferentes teorias relacionadas à recuperação da Informação. Compreende uma informação tida como suplementar, ou seja, que serve ao indivíduo como algo que o ajude a conhecer melhor o mundo no qual vive. Apóia-se nas Ciências Cognitivas, Sociologia e Inteligência Artificial, questões que foram abordadas por Ellis, Shank, Abelson, Johnson-Laird, entre outros.

A Figura 1 retrata as três abordagens acima descritas.

ABORDAGENS	
	Tradicional
	Orientada ao usuário
	Cognitiva

Figura 1 - Abordagens da Recuperação da Informação

Apesar de as três abordagens tratarem da RI, a primeira (a abordagem tradicional) diz mais respeito ao corpus conceitual no qual esta pesquisa sobre a **PL** se concentra. Os principais motivos que sustentam essa posição são: primeiro, o público–alvo e também os conteúdos da **PL**, que estão direcionados a uma comunidade delimitada: a científica e a tecnológica; segundo, a **PL**, como um sistema de RI, apóia-se nas representações dos registros (currículos) do seu acervo. O sucesso da recuperação da **PL** depende do nível de coincidência – aqui entendido como relevância - entre uma dada estratégia de busca[19] e as representações dos registros do acervo.

Tão relevante quanto as abordagens são os modelos de RI. Segundo Baeza-Yates e Ribeiro-Neto (1999), há duas formas de buscar informações e nelas modelos de recuperação. Uma das formas, a de **navegação**[20]**,** baseia-se em recursos navegacionais usando o hipertexto (forma não linear), roteiros estruturados (hierárquicos) e planos (bi–dimensionais). Para esta pesquisa não é relevante estudar esta forma, pois a **PL** não está fundamentada em tais recursos. A outra forma de busca (a mais usual) é a **Recuperação *ad hoc* e de Filtragem,** que se subdividem em Modelos Clássicos e Modelos Estruturados.

[19] Considera-se estratégia de busca o conjunto de termos (palavras) explicitado pelo usuário para efetuar uma busca em sistema de recuperação da informação. Alguns sistemas podem aceitar comandos com operadores booleanos (and, not, or), ou outros especiais como caracteres curingas (?, *).

[20] Originalmente Browsing.

Nos modelos clássicos, um documento é representado por palavras-chave que representam a temática do documento e ainda sintetizam o seu conteúdo. Nos modelos estruturados, além da representação por palavras-chave, são incluídas informações sobre a estrutura do texto, que permitem fazer buscas através do coeficiente de proximidade entre palavras, parágrafos de documento, formatações no texto, etc.

Baseando-se em Baeza-Yates & Ribeiro-Neto (1999) e Souza (2006), serão detalhados os três modelos clássicos de **RI**, que são: o modelo booleano, o modelo vetorial e o modelo probabilístico, conforme sintetizado na Figura 2:

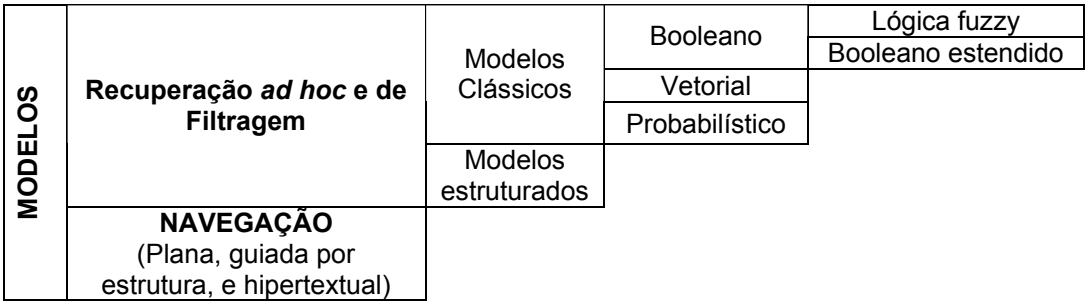

Figura 2 - Modelos de Recuperação da Informação

a) Modelo Booleano: baseia-se na teoria dos conjuntos, e não é visto como um dos modelos mais eficazes, porém, destaca-se por sua simplicidade e por ter sido amplamente empregado nos sistemas bibliográficos comerciais, particularmente antes do advento da internet. Nas buscas, recupera documentos (mais precisamente suas representações) coincidentes com a estratégia formulada (através de termos ou palavras-chave) pelo usuário. A coincidência ocorre em um nível de correspondência binária, o que significa que a grafia do(s) termo(s) da estratégia de busca e a representação precisam ser idênticas.

Há operadores (AND, OR e NOT) que criam relacionamentos, isto é, possibilitam ao usuário formular operações lógicas com as palavras-chave para alcançar resultados mais refinados. Sua maior desvantagem, além de trabalhar de forma binária, no qual os documentos são analisados de forma dual (relevante ou não relevante), é não prever formas de ordenação dos resultados por grau de relevância (SOUZA, 2006). Duas correntes mais aperfeiçoadas do

modelo booleano são a Lógica Fuzzy (também conhecida como difusa ou nebulosa) e o Booleano estendido.

De acordo com Ferneda (2004), a lógica fuzzy (ou lógica difusa) busca lidar sistematicamente com a diversidade, a incerteza e as verdades parciais dos fenômenos da natureza. Para tanto, é ampliada a capacidade de representação das palavras-chave estipulando-se para cada termo contido na base de dados, níveis de relações semânticas com documentos.

Para representar um documento, a lógica fuzzy adota uma função que atribui valores, que serão os pesos de cada termo para o documento. Os pesos associados a um termo expressarão o quanto é significativo ou não na descrição do conteúdo do documento.

> A qualidade da recuperação depende em grande parte da função adotada para calcular os pesos dos termos de indexação. Geralmente esta função baseia-se no cálculo da freqüência de ocorrência dos termos em todo o texto, e fornece uma representação estática do documento (FERNEDA, 2004, p.46)

Na lógica Fuzzy, a adoção de instrumentos como o tesauro poderia contribuir para a indicação da pertinência ou não de um termo a um determinado conjunto semântico.

O modelo Booleano estendido teve pouca utilização, mas serviu como núcleo do modelo vetorial. A intenção do Booleano estendido era superar o problema das decisões binárias do booleano clássico, através da atribuição de pesos aos termos.

Ressalta-se que a **PL** adota o modelo booleano, com isso as buscas no sistema seguem um princípio simples: se na busca for definido um termo que sintaticamente, não corresponda a nenhum outro da base da **PL**, não haverá currículo recuperado. A busca pode ser refinada com o uso de operadores booleanos como o “AND”, o “OR”, e o “NOT” e, para tanto, é necessário combinar, no mínimo, dois termos. O exemplo abaixo demonstra o uso de operador booleano:

- Um usuário define o termo **THESAURUS** para efetuar uma busca na **PL**. Se não constar na base do sistema nenhuma referência ao termo THESAURUS, o resultado da busca não trará nenhum registro. Numa segunda situação, se o usuário efetuar uma busca por um termo genérico como CONHECIMENTO, muitos resultados serão encontrados. Porém, se o usuário fizer a seguinte combinação: THESAURUS AND CONHECIMENTO, apesar da grande

quantidade que o termo CONHECIMENTO recuperaria, a combinação com o termo THESAURUS devolveria um resultado com nenhuma ocorrência.

Como a indexação dos currículos da **PL** é feita pelos usuários, ao incluírem as palavras-chave, essas passam a valer como termos representativos dos seus respectivos currículos. No exemplo observa-se que há deficiência no modelo booleano para operar com simples relações de equivalência, pois a grafia THESAURUS impediria de se buscar currículos representados por TESAURO. Esta situação seria contornável se um recurso de equivalências fosse implementado na **PL**.

Devido às deficiências do modelo booleano, no que diz respeito a determinados aspectos da recuperação da informação, é que se desenvolveram alternativas de modelos como o vetorial e o probabilístico.

b) Modelo Vetorial: O modelo vetorial possibilita recuperar documentos que respondam parcialmente a uma estratégia de busca e, para realizar esta tarefa, associa pesos tanto aos termos de indexação como aos termos da estratégia de busca. Ferneda (2003, p.27-28) explica que esses pesos servem para calcular o grau de similaridade entre a expressão de busca formulada pelo usuário e cada um dos documentos do acervo e oferece no resultado "um conjunto de documentos ordenados pelo grau de similaridade de cada documento em relação à expressão de busca".

C) Modelo Probabilístico: este modelo valoriza a interação do usuário com o sistema, fundamentando-se no seguinte princípio: há um conjunto recuperável (e ideal) de documentos que responde a cada busca realizada no sistema. Para identificar esse conjunto ideal, definem-se arbitrariamente conjuntos de documentos que servirão para medir o *feedback* dos usuários com relação a estes conjuntos em determinadas buscas. Analisando-se a interação do usuário, identificam-se quais os documentos mais relevantes em situações específicas de buscas.

Os três modelos (booleano, vetorial e probabilístico) não se excluem e não foram criados de forma isolada. Eles foram desenvolvidos no intuito de modelar relações entre objetos a partir de fundamentos lógicos. A priori, descrever simbolicamente uma realidade é uma tarefa comum a disciplinas como a Matemática e a Física. Entretanto, trabalhar com representações

simbólicas culturalmente produzidas - como a informação - não se assemelha a modelar uma realidade natural.

Nesta condição, entre o natural e o cultural, foram estabelecidos os princípios da RI em meio eletrônico, que por um lado necessita modelos racionais que forneçam algoritmos computacionais para processar rotinas a partir de dados contidos numa base. Por outro lado, os dados que constituem esta base são representações de conteúdos semânticos, dotados de significado, de sentido e de contextos de produção e uso.

Para sistematizar a discussão sobre RI, apresentam-se as abordagens e modelos (Figura 3 e Figura 4) indicando-se (nas células na cor cinza) as categorias nas quais a **PL** se insere:

ABORDAGENS	Tradicional
	Orientada ao usuário
	Cognitiva

Figura 3 - Abordagens da Recuperação da Informação da Plataforma Lattes

MODELOS	**Recuperação *ad hoc* e de Filtragem**	Modelos Clássicos	Booleano	Lógica fuzzy
				Booleano estendido
			Vetorial	
			Probabilístico	
		Modelos estruturados		
	NAVEGAÇÃO (Plana, guiada por estrutura, e hipertextual)			

Figura 4 - Modelos de Recuperação da Informação da Plataforma Lattes

3.2.1.1 Sistemas de Recuperação da Informação

Além de apresentar abordagens e modelos, é apropriado ampliar a discussão sobre RI a um universo maior, o dos Sistemas de Recuperação da Informação (SRI).

A definição de SRI não é simples. Para Souza (2006) a dificuldade resulta da ambigüidade dos conceitos de sistema e de informação. Não é necessário rediscutir tais conceitos, por isso será enfocada somente a função dos SRI como intermediários das necessidades informacionais do usuário. Souza (2006, p.162) baseia-se em Lancaster e Warner para afirmar que

> os SRI são a interface entre uma coleção de recursos de informação, em meio impresso ou não, e uma população de usuários; e desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle desses; e distribuição e disseminação aos usuários. Essa visão é abrangente, e inclui tarefas que são desempenhadas em conjunto com atores humanos.

Os SRI não informam o usuário, indicam somente a existência de documentos pertinentes às suas necessidades informacionais (e descrevem esses documentos). Além disso, os SRI não respondem às necessidades informacionais dos indivíduos, mas apenas às representações de suas necessidades. Desta maneira, um indivíduo, além de reconhecer a sua necessidade, precisa, presumivelmente, estar estimulado a seguir alguns passos para satisfazê-la (LANCASTER, 1979).

Os SRI tiveram, basicamente, duas linhas históricas de desenvolvimento: uma, originada nos grandes sistemas[21]que operavam com termos extraídos de um vocabulário controlado e atribuídos aos documentos por pessoas; e outra, originária da área jurídica, que se distinguia por inserir documentos completos (como leis) em formato eletrônico, além de utilizar computadores nas buscas por palavras nesses textos, cujos conteúdos eram representados sem intervenção humana (LANCASTER, 2004).

Já na década de 1960, Kent (1972, p.24-26) afirmava que um SRI, informatizado ou não, seguia sete operações chamadas por ele de “unitárias”. É válido apresentá-las:

a) **análise** – leitura do documento para seleção dos itens mais relevantes que mereçam o esforço em torná-los reconhecíveis pelos usuários do sistema. É uma análise que antecede a representação;

b) **controle de vocabulário e rubrica de assunto** – padronização dos assuntos e estabelecimento de relações entre eles, na linguagem do sistema ;

c) **registro dos resultados da análise em um instrumento passível de pesquisa** – diz respeito ao registro de fato no sistema, ou seja, a inserção de dados;

d) **armazenagem de registros ou documentos-fonte** – inserção no acervo de um documento que foi previamente representado;

[21] Instituições como a National Library of Medicine (NLM), o Departament of Defense (DOD), e a National Aeronautics and Space Administration (NASA)

e) **análise de questões e desenvolvimento de uma estratégia de pesquisa** – inclui a formulação de uma estratégia de busca a partir dos recursos oferecidos pelo sistema;

f) **condução da pesquisa** – etapa final que precede a resposta que o sistema fornecerá como resultado da busca;

g) **exposição dos resultados da pesquisa** – resultado propriamente dito da busca.

Não é viável confrontar perfeitamente as sete operações de Kent com o atual contexto dos SRI, pois mudanças ocorreram em mais de 40 anos da publicação das idéias desse autor. Entretanto, buscou-se fazer uma analogia entre suas operações e o funcionamento da **PL**.

Na **PL,** a **Análise** é elaborada pelo autor do currículo e cabe a ele preencher os campos que descrevem suas atividades. Por outro lado, na **PL** há um processo de indexação automático que considera todas as palavras registradas no currículo. Assim, qualquer trecho do currículo torna-se uma representação para as buscas. Desse modo, o item **controle de vocabulário e rubrica de assunto** é o processo mais fragilizado da **PL,** devido à ausência de controle no sistema. O único "controle" de termos que existe está nos campos da categoria Sem Autonomia, que será explorada na seção 4 (Análises da **PL**).

A etapa **registro dos resultados da análise em um instrumento passível de pesquisa** seria, na **PL**, a integração do processo de preenchimento pelo autor do currículo e da indexação automática, para constituir um índice. É a partir deste índice que as buscas serão efetivadas.

A etapa de **armazenagem de registros ou documentos-fonte** na **PL** é o próprio processo de criação/manutenção dos currículos, já que cada currículo é um documento do sistema e o ato de criá-lo ou atualizá-lo implica na criação ou atualização de um documento.

As etapas de **análise de questões e desenvolvimento de uma estratégia de pesquisa** e a de **condução da pesquisa** dizem respeito às buscas no sistema. Na **PL**, esse processo pode ser feito nas interfaces de busca simples ou avançada. A diferença entre as duas está na quantidade de campos oferecidos aos usuários para montar uma estratégia de busca. Na interface de busca simples é possível inserir somente palavras para efetuar buscas por nome e por assunto (e limitar a busca para recuperar apenas

currículos de doutores). Na interface de busca avançada o usuário pode formular sua estratégia usando operadores booleanos e ainda refinar os resultados a partir de campos[22] oferecidos na interface de busca.

Após definida a estratégia, a condução de pesquisa é efetivada enviando-se os dados. Em seguida, o usuário terá a **exposição dos resultados da pesquisa**, que podem ser mostrados por ordem alfabética de nomes ou através de um "score" baseado em critérios da produção dos pesquisadores.

As TICs progrediram bastante desde a publicação do texto de Kent até os dias atuais; porém, a estrutura básica de funcionamento de um SRI tem ainda elementos em comum com os modelos antigos. É visível que ocorreram mudanças nos procedimentos adotados em cada etapa descrita por Kent, contudo as finalidades dos processos que envolvem o funcionamento de um SRI continuaram essencialmente as mesmas.

A seguir, há dois modelos de atividades comuns em SRI: um elaborado por Lancaster (Figura 5) e outro por Baeza-Yates e Ribeiro-Neto (Figura 6). Tais modelos foram criados em períodos distintos (respectivamente nas décadas de 1960 e 1990), mas percebe-se em ambos atividades que têm como núcleo uma fase específica para a representação do conteúdo do documento, denominada por Lancaster "análise conceitual" e por Baeza- Yates e Ribeiro Neto "indexação".

[22] Os campos que podem ser utilizados para refinar a busca na PL são: pesquisadores do CNPq, bolsistas do CNPq, formação acadêmica, nível do curso de pós-graduação onde é docente, área de atuação, atividade de orientação, idioma, áreas ou setores da produção em C&T, atividade profissional e presença no Diretório de Grupos de Pesquisa.

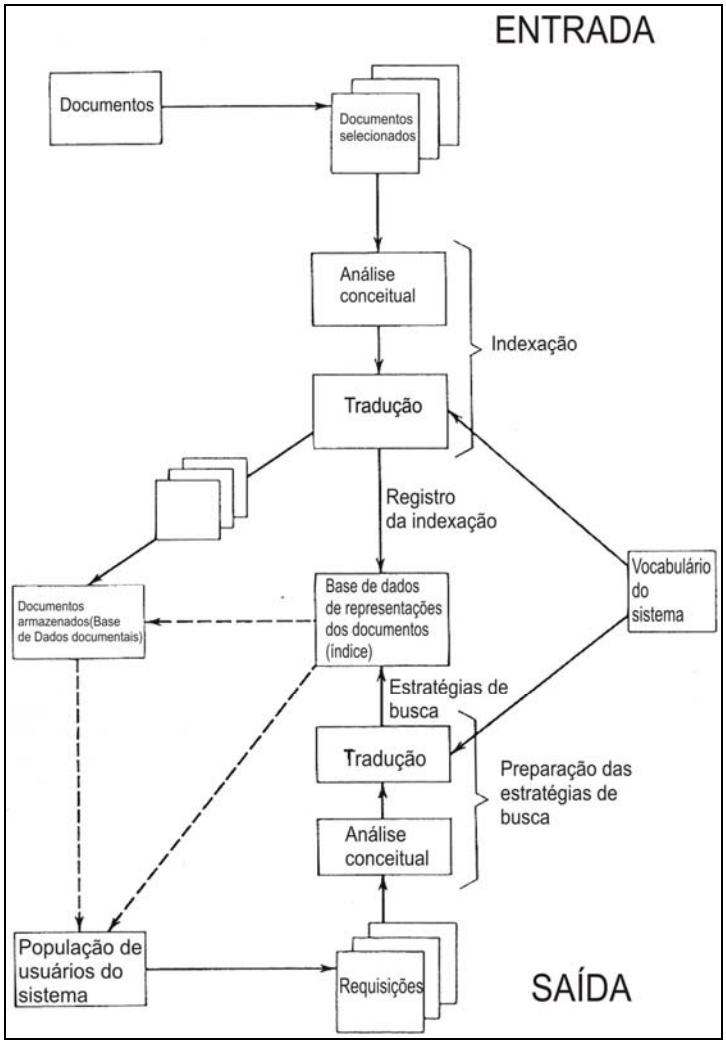

Figura 5 - Atividades freqüentes em SRI (LANCASTER, 1979)

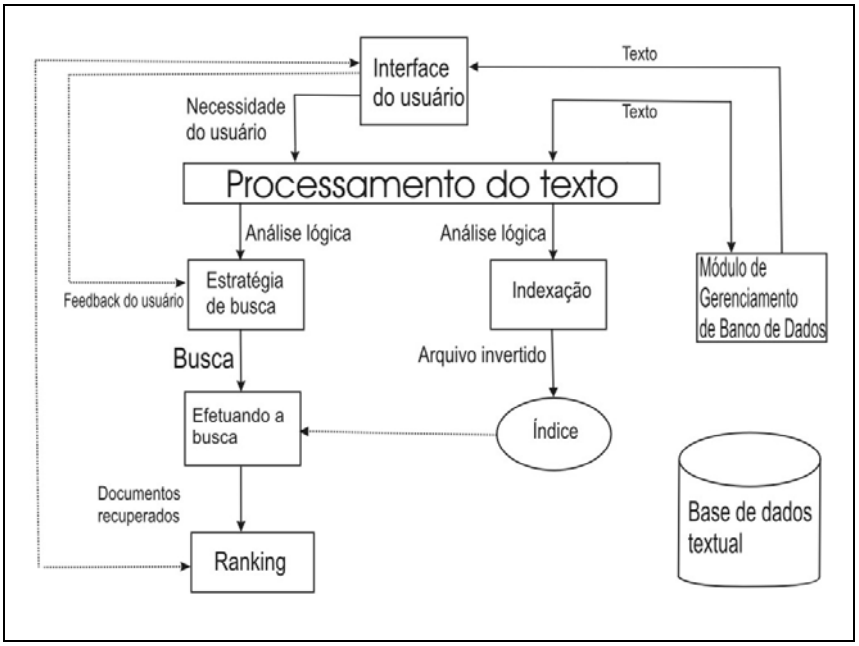

Figura 6 – O processo de recuperação da informação (BAEZA-YATES E RIBEIRO-NETO, 1999)

Na etapa da indexação percebe-se que Lancaster está mais atento às representações temáticas (ou conceituais) do documento, enquanto Baeza-Yates e Ribeiro-Neto não distinguem as representações temáticas das descritivas. Um provável motivo para tal diferença é que no esquema proposto por Lancaster - na década de 1960 - incluía-se a participação humana na indexação e uso de vocabulários controlados, enquanto que no esquema de Baeza-Yates e Ribeiro Neto os processos modelados matematicamente em algoritmos nem sempre contam com a presença de pessoas.

Nos últimos anos, a indexação automática nos SRI vem crescendo, e os recursos baseados em busca textual têm se consolidado e até mesmo despontado como uma tendência, ao menos é o que se percebe em sistemas de buscas genéricas na Internet. A opinião de Lancaster (2004, p.252) é que uma

> [...] distinção entre os sistemas baseados essencialmente em vocabulários controlados e registros de indexação criados por seres humanos [...] e os sistemas baseados em buscas no texto tem se tornado cada vez mais difusa com o passar dos anos. [...] Os sítios da rede da Internet consistem majoritariamente em texto, de modo que uma verdade indubitável é que as buscas em textos superam hoje grandemente as buscas que envolvem vocabulários controlados.

A **PL** possui uma peculiaridade, comparando-se a outros SRI. Seu acervo é formado por currículos que são simultaneamente conteúdo e representação do seu conteúdo, pois os registros são automaticamente indexados para posterior recuperação. Assim, são criados índices - arquivos invertidos - a partir dos currículos, constituindo uma base de palavras que indicarão em quais registros constam aquela determinada palavra ou expressão, ou seja, a PL funciona como sítios da Internet, os currículos sendo assimilados a textos mas contam com um refinamento: a inclusão de palavras-chave pelos “autores”. Na **PL**, a participação humana na etapa de representação é condição básica, assim como a interferência humana é muito importante em qualquer sistema baseado nos modelos de auto-arquivamento, em que os próprios usuários inserem e muitas vezes categorizam e descrevem os novos documentos.

Outra discussão pertinente aos SRI diz respeito à avaliação dos sistemas. Senko destaca que “sem dúvida a avaliação é a área mais problemática dos SRI” (VAN RIJSBERGEN, 1979, p.6). Uma forma de avaliar

os SRI é através da *relevância*[23], que em testes controlados de laboratórios se demonstra eficaz, contudo Cuadra e Katter (citados por Van Rijsbergen, 1979), perceberam que variáveis externas (usualmente não controláveis em laboratórios) podem distorcer os resultados. Para Van Rijsbergen a efetividade da recuperação é mais bem avaliada através dos coeficientes de *precisão* e *revocação*.

De acordo com Lancaster (2004), o coeficiente[24] de precisão (cp) reflete a proporção entre o número de itens que o usuário deve analisar para selecionar aqueles que serão relevantes. A seleção é realizada a partir dos resultados obtidos em uma busca feita no sistema, e pode ser assim representada:

$$cp = \frac{\text{ítens considerados relevantes}}{\text{total de itens recuperados}}$$

Para exemplificar, supõe-se que fosse feita uma busca na **PL** usando o termo *BOOLEANO* que recuperou 18 registros. Se houvessem critérios que indicassem que entre os registros recuperados, 3 fossem relevantes, o coeficiente de precisão seria $\frac{3}{18}$ (ou 16,6%).

Já o coeficiente de revocação (cr) é o número de documentos relevantes recuperados pelo sistema, dividido pelo número total de registros relevantes existentes no sistema. Pode ser representado da seguinte forma:

$$cr = \frac{\text{documentos relevantes recuperados}}{\text{todos os registros relevantes do sistema}}$$

Neste caso, seria necessário saber quantos currículos condizentes com a temática BOOLEANO existem na **PL**. Não seria simples, pois podem haver, por exemplo, vários currículos relacionados ao assunto *RECUPERAÇÃO DA INFORMAÇÃO* que sejam relevantes para quem estiver interessado em "operadores booleanos". Assim, fez-se uma busca usando a expressão RECUPERAÇÃO DA INFORMAÇÃO, em que foram recuperados 684 registros. O número total de documentos relevantes no sistema dependeria de uma

[23] Segundo Lancaster (2004, p.14) a relevância é "a relação entre um documento e uma necessidade de informação ou entre um documento e um enunciado de necessidade de informação (uma consulta)".
[24] Autores como Robredo (2005, p.200) preferem índice de precisão e índice de revocação (ou exaustividade) a coeficiente.

avaliação por um “juiz”, ou pela pessoa que efetuou a busca: o conceito de “relevância” remete sempre a uma avaliação subjetiva, ou contextualizada no tempo e no espaço.

Por deficiências desta natureza, alguns autores vêem pouca aplicabilidade no coeficiente de revocação. É o caso de Boccato e Fujita (2006, p.270) que, ao revisarem estudos a respeito de avaliação de linguagens documentárias, perceberam que

> para obter-se o número de referências relevantes existentes no sistema talvez fosse preciso a realização de uma pesquisa muito genérica sobre um determinado assunto; porém, este não seria necessariamente o intuito normal de um usuário (BOCCATO E FUJITA, 2006, p.270)

Por fim, outra particularidade inerente à recuperação da informação é que há relação direta e recíproca entre os coeficientes de precisão e revocação: a melhoria em um dos coeficientes, em geral, implica em perda para o outro (LANCASTER, 2004, p.4). Para melhor entendimento, utilizaremos o exemplo de Lancaster (2004, p.3-4):

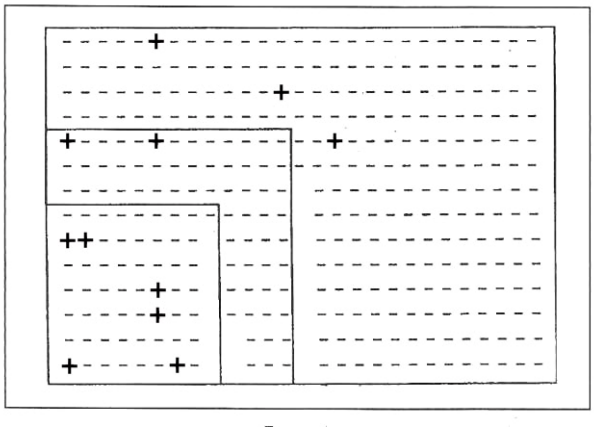

Figura 7 - O problema da recuperação de itens pertinentes de uma base de dados. (LANCASTER, 2004, p. 3)

O retângulo maior representa uma base de dados, os itens com o sinal de adição (+) são aqueles considerados úteis para uma determinada busca e os itens com o sinal de subtração (–) são os não considerados úteis. O retângulo menor representa uma busca realizada na base de dados, que recuperou 57 itens - seis foram úteis e 51 inúteis. Assim, a relação entre úteis e inúteis ($\frac{6}{57}$ ou 10%) é o coeficiente de precisão. E o índice de revocação expressaria a extensão dos itens úteis que, neste exemplo, seria de 6/11 ou 54%.

O segundo maior retângulo interno indica uma busca mais genérica, com uma taxa de revocação que subiu para $\frac{8}{11}$ ou 73%; por outro lado, a taxa de precisão reduziu para $\frac{8}{112}$ ou 7%.

3.3 ABORDAGENS TRADICIONAIS PARA A ORGANIZAÇÃO DA INFORMAÇÃO

Recursos para organizar a informação buscam organizar, gerenciar e recuperar a informação. Tais recursos se fazem presentes em muitas áreas do conhecimento humano, seja em estruturas mais simples como mais complexas. Para Tristão et al (2004), no âmbito da organização da informação, esses recursos abrangem: classificação, tesauro, ontologia, glossários e dicionários, específicos a cada área e, em sua maioria, ligados a bibliotecas e outras organizações de gerenciamento da informação. Além dos recursos citados pelas autoras, acrescenta-se os mapas conceituais. Para os interesses deste trabalho serão apresentados a classificação, o tesauro e as ontologias.

Classificar é ordenar, agrupar, organizar, segundo características em comum. Além disso, o ato de classificar serve a propósitos que definirão o grau de complexidade das classificações. Para Svenonius (2001), a organização pode assumir várias formas, contudo sua forma sistemática[25] é a classificação, que agrupa coisas através de semelhanças, a partir de um ou mais atributos. Diemer, citado por Pombo (2007), diz que há quatro grandes orientações à classificação:

- uma orientação ontológica (classificação dos seres);
- uma orientação gnosiológica (classificação das ciências);
- uma orientação biblioteconômica (classificação dos livros);
- uma orientação informacional (classificação das informações).

A orientação ontológica (classificação dos seres) atende a problemas da classificação nas ciências, que teve origens com Aristóteles e, atualmente, interessa aos lógicos e cientistas de áreas como a Biologia, a Geologia, a Cosmologia, e a Antropologia. A orientação gnosiológica (classificação dos

[25] No original a autora citou prototypical form.

saberes) diz respeito ao problema da classificação das ciências, de maior interesse dos filósofos e daqueles que refletem a Ciência e sua produção.

Mas a orientação biblioteconômica e informacional (classificação dos livros e das informações) tem maior relação com esta pesquisa e, de acordo com Pombo (2007, p.3), “corresponde à constituição de uma ciência da classificação, isto é, de um novo domínio científico que tem por tarefa o estudo de todos os possíveis sistemas de classificação”.

Segundo a citada autora, a diferença entre as classificações das ciências e as classificações informacionais e biblioteconômicas reside na maneira especulativa das primeiras em contraste com a predominância mais funcional das segundas. Enquanto as primeiras são universais, genéricas e não se apegam às minúcias de classificação de domínios restritos, as segundas são especializadas.

No bem elaborado histórico das classificações biblioteconômicas, Lima (2004) percorre o trajeto de Alexandria aos dias atuais e explora a questão da incapacidade dos antigos sistemas de classificação para lidar com os crescentes volumes de informação. A mencionada incapacidade demandou o aperfeiçoamento das formas de classificação da informação. Nesse ínterim é que foram desenvolvidas as Linguagens Documentárias (LD), que segundo Cintra et al (2002), são linguagens construídas para a indexação, armazenamento e recuperação da informação e são destinadas à “tradução” dos conteúdos dos documentos. As LDs mais conhecidas são os sistemas de classificação bibliográficos (Classificação de Harris, a Classificação Decimal de Dewey (CDD), a Classificação Decimal Universal (CDU), a Classificação da Biblioteca do Congresso (LC), a Classificação de Dois Pontos) e os tesauros, que podem ser facetados ou não facetados.

Para Iyer (1995) os esquemas não-facetados são enumerativos e arrolam cada um dos possíveis elementos, combinações e conjuntos de assuntos existentes em uma área, enquanto os facetados se baseiam numa combinação de pequenos grupos conceituais melhores que listas, e sua base é a decomposição de conceitos até todas as possíveis características.

As LDs ganharam força no período entre as décadas de 1950 e 1960, quando as formas de armazenar e recuperar as informações não acompanharam o crescimento do conhecimento científico e tecnológico, assim

perdeu força “[...] a perspectiva preferencial de recuperação bibliográfica e normalização classificatória e descritiva, buscando-se a construção de linguagens próprias.” (CINTRA et al, 2002, p.33). Na definição de Gardin citado por Cintra et al (2002, p.35) uma LD

> é um conjunto de termos, providos ou não de regras sintáticas, utilizadas para representar conteúdos de documentos técnico-científicos com fins de classificação ou busca retrospectiva de informações.

Cintra et al (2002, p.34-35) destacam as seguintes características das LD:

- através delas pode-se representar, de forma sintética, informações materializadas em textos;
- apesar de, assim como a linguagem natural, ser um sistema simbólico instituído para fins de comunicação, a LD é restrita a contextos documentários, objetivando tornar possível a comunicação usuário-sistema de informação;
- o sistema de relações de uma LD, bem como seus mecanismos de relações, quando comparados aos da linguagem natural, são precários. Os elementos das LD são selecionados de determinados universos nos quais se constrói seu sistema de relações, que só poderá ser utilizado se houver regras explícitas. Por isso se diz que as LD são linguagens construídas;
- são instrumentos intermediários, através dos quais se busca unir a pergunta do usuário às unidades informacionais do sistema, numa linguagem que é própria do sistema.

Mencionou-se anteriormente que um dos problemas centrais desta pesquisa é a ausência de controle da **PL** quanto à alimentação de suas bases de dados. Adianta-se que uma das particularidades do sistema é que há somente um pequeno nível de controle para preenchimento dos campos categorizados como Sem Autonomia, mas para outros não, ou seja, o sistema combina o uso da linguagem natural e documentária, como veremos em detalhe na seção 4.

Sobre o uso de Linguagem Natural combinado com a Linguagem controlada, há uma interessante revisão de literatura feita por Lopes (2002, p.51) que, entre outros itens, destaca:

a) Linguagem Natural

Vantagens: registro imediato da informação sem consulta a uma linguagem de controle; a busca não requer treinamentos específicos no uso de uma linguagem de controle; os termos de entrada de dados são extraídos diretamente dos documentos; indexadores e usuários têm acesso aos mesmos termos.

Desvantagens: maior esforço intelectual para identificar os sinônimos, as grafias alternativas, os homônimos, etc, na busca; alta incidência de desentendimento entre os termos da busca e os do sistema; a estratégia de busca requer todos os principais conceitos e sinônimos.

b) Linguagem Controlada

- Vantagens: problemas de comunicação entre indexadores e usuários são minimizados graças ao controle total do vocabulário de indexação; indexadores atribuem melhor os conceitos dos documentos utilizando um tesauro; um vocabulário controlado pode proporcionar alta precisão nos resultados, ampliando a confiança do usuário; as relações hierárquicas e as remissivas do vocabulário controlado auxiliam o indexador e usuários na seleção de conceitos.

- Desvantagens: alto custo na produção e manutenção da base de dados e necessidade de manter pessoal especializado na atualização do tesauro; o vocabulário controlado desatualizado pode não se adequar aos objetivos do produtor da base; um vocabulário controlado poderá se distanciar dos conceitos adequados para a representação das necessidades de informação dos usuários; possibilidade de falsos resultados por conta da desatualização do vocabulário controlado.

Assim, ainda que a adoção de linguagem natural combinada com recursos de linguagem controlada em SICT possa oferecer bons resultados, o sucesso dela é mais perceptível em contextos como os das bases de dados comerciais, planejadas previamente com o propósito de seus sistemas servirem como recurso para a RI.

Na avaliação de Lopes (2002) percebe-se que o tesauro recebe uma atenção especial nas discussões relativas à organização da informação. O

tesauro[26] é uma das mais importantes modalidades de LD e, assim como outras, surge como resposta à ineficiência dos recursos de organização da informação que não atendem às demandas impostas pelo ambiente da produção de documentos especializados. “Era preciso trabalhar com vocabulário mais específico e com uma estrutura mais depurada do que aquela presente nos cabeçalhos de assunto (remissivas e referências cruzadas tipo ver e ver também)” (DODEBEI, 2002, p. 66).

O objetivo maior do tesauro, de acordo com Cintra et al (2002), é o controle terminológico, que pode ser alcançado com modificadores que contextualizam o sentido pretendido, e com definições e notas de escopo que evitam duas ocorrências: a da **polissemia** (dependendo do contexto uma palavra pode comportar mais de um significado) e a da **homonímia** (diferentes objetos designados pela mesma palavra). Essas ocorrências são comuns na linguagem natural, porém devem ser evitadas numa linguagem controlada como o tesauro, a fim de evitar a ambigüidade (mais de uma interpretação no processo da comunicação lingüística).

No tesauro busca-se a monossemia dos termos, para que uma única forma significante corresponda a um único significado. As redes relacionais são igualmente necessárias, pois estabelecem a posição dos termos com relação a outros termos do sistema, conduzindo a um maior controle terminológico e permitindo que nenhuma unidade presente numa LD não esteja relacionada a uma outra unidade.

A partir das relações entre os termos de um tesauro forma-se uma rede paradigmática. Essas relações podem ser expressas pelo sistema nocional de forma hierárquica ou associativa. As relações hierárquicas compreendem as relações genéricas, específicas e partitivas, por elas determinam-se as relações entre o gênero e a espécie, ou entre o todo e suas partes. As relações hierárquicas se expressam nos níveis de superordenação e subordinação de um termo em relação ao outro, e se estiverem em níveis idênticos de subordinação, tornam-se coordenados.

O tesauro pode auxiliar o usuário nas buscas informacionais, como ajudar o indexador durante o processo de classificação. Sua estrutura de

[26] Aqui trataremos do tesauro documentário

termos e suas relações auxiliam a encontrar o melhor termo ou termos que representem um assunto. Moreira, Alvarenga e Oliveira (2004) consideram que o tesauro,

> é um componente muito importante num sistema de recuperação por cumprir o papel de: determinar quais termos podem ser usados no sistema; determinar quais termos podem ser usados na busca para que esta tenha um resultado satisfatório; e permitir a introdução de novos termos em sua estrutura de termos e relações de modo a aproximar a linguagem do usuário à do sistema e realizar alterações de sentidos dos termos existentes.

Concorda-se com a opinião das autoras, complementando-se que num ambiente como o da C&T, tanto na produção como no uso de sistemas eletrônicos de informação, o tesauro é importante, porém sua elaboração e manutenção representam um custo relativamente alto. Infelizmente, SICTs abertos como a **PL** são criados numa perspectiva quantitativa de composição de estoques de informação, uma vez que tal modelo segue uma lógica econômica na produção da informação científica. Manter uma posição dicotômica, ou seja, desenvolver e manter SICTs mais consistentes ou mais econômicos, não favorece as formas de tratamento dos estoques de ICT, pois a perspectiva de uso dos sistemas híbridos contempla a combinação de recursos tradicionais (como os tesauros) com novos instrumentos (como as ontologias). A partir da combinação de recursos é provável que sistemas menos onerosos e mais consistentes sejam desenvolvidos.

A respeito da lógica econômica, Bolaño, Kobashi e Santos (2006) discutem a produção científica certificada que, apesar de desviar do foco desta tese, coincide em um ponto: o tratamento e organização da ICT já não se restringem a uma orientação qualitativa que almeja resolver problemas da recuperação da informação. Foram fortalecidos os recursos técnicos baseados no uso das TICs que buscam aperfeiçoar os recursos informacionais por meio de tratamentos automatizados. Estes, por sua vez, buscam lidar com volumes cada vez maiores de informação por um custo cada vez menor, incorporando um aspecto quantitativo ao qualitativo.

3.4 ORGANIZAÇÃO DA INFORMAÇÃO EM MEIO ELETRÔNICO

O uso das TICs voltado à informação em meio eletrônico teve seus primeiros passos nas décadas de 1940 e 1950. Nesse período, personalidades destacaram-se por discutirem as formas de lidar com o conhecimento humano registrado e uma dessas pessoas foi Vannevar Bush, um autor bastante citado na área da Ciência da Informação graças à publicação do artigo “As we may think” (BUSH, 1945). Esse artigo, comumente citado (e provavelmente pouco lido), baseia-se, genericamente, em duas vertentes: uma volumosa produção da informação no período pós-guerras mundiais e previsões de novos instrumentos para lidar com tais estoques.

Allen Kent (1972) foi outra personalidade importante, por ter analisado e descrito conceitos e instrumentos relacionados à informação, abordando temáticas como a recuperação, classificação, indexação e até mesmo o gerenciamento de unidades de informação como as bibliotecas.

Esse autor demonstrou conhecer bem os fundamentos conceituais da informação adotados por áreas como a Documentação, Biblioteconomia e Lingüística, contudo a sua visão aponta um forte viés para uma noção matemática da informação. Também é perceptível, nas palavras de Kent (1972, p. 240), uma opinião desfavorável às LD:

> Desenvolveu-se um certo número de linguagens artificiais procurando evitar ambigüidades tanto em relação ao significado das palavras existentes em seus vocabulários como nos modelos sintáticos empregados para representar as relações entre as palavras. Essas linguagens artificiais, embora as regras do seu emprego não sejam ambíguas, pagam um preço por tal vantagem, pois perdem a expressividade e riqueza, como também a flexibilidade da língua natural.

Em outro trecho, ao debater a função das palavras, da linguagem e do significado nos sistemas de recuperação, Kent (1972, p.241) esclarece que

> de um ponto de vista prático, o significado não tem tanta importância num sistema de recuperação da informação, a não ser na medida em que auxilia um cliente a localizar o registro que deseja.

Com tal afirmativa Kent subestima que as representações lingüísticas são construídas justamente a partir de contextos e significados. Para ele, não havia importância se o analista do documento (pessoa responsável por introduzir os dados no sistema) desconhecesse o significado do termo a ser

inserido. O que importava – tanto para quem alimentava o sistema como para quem o utilizasse - era a grafia correta da palavra.

Contundo, Kent (1972) estava ciente das vantagens e desvantagens proporcionadas por instrumentos como o tesauro. Mas, conforme o que já foi dito no início deste trabalho – usando palavras de Foskett -, a década de 1970 foi marcada pela ênfase no processamento automático da informação, negligenciando-se o significado que a informação comporta como comunicadora do conhecimento.

As soluções para lidar com a "explosão documental" buscaram um tratamento da informação orientado por uma linha mais racional e econômica, que reduziu o tratamento da informação a operações matemáticas. Ganha-se em parte, ao processar volumes de informação com velocidade jamais alcançada por humanos, contudo, qualitativamente, o tratamento da informação desprovido de significado pode resultar em recursos informacionais com limitações de natureza semântica. Entre algumas limitações destacam-se: dificuldades para contextualizar domínios específicos de conhecimento, inaptidão para lidar com sinonímias e homonímias e ausência de recursos que estabeleçam relações nocionais entre conteúdos afins.

Essas limitações podem ser percebidas em sistemas genéricos de busca, como os da Internet. Para ilustrar o problema, fez-se uma busca no Google[27] usando o termo "*coração*"[28]. Os primeiros resultados foram:

1 – sítio de uma empresa de publicidade contendo uma animação sobre o amor (coração no sentido metafórico de sentimentos);

2 - página com informações básicas sobre o corpo humano (o coração como órgão);

3 - páginas institucionais do: HCOR - Hospital do Coração, Instituto do Coração (InCor) do Hospital das Clínicas da Faculdade de Medicina da USP, e da Universidade do Sagrado Coração;

4 - páginas pessoais com mensagens de auto-ajuda, poesias, epígrafes, etc.

Esse foi um exemplo ilustrativo sem pretensões metodológicas, mas qualquer outra busca feita em um sistema genérico de recuperação da informação tende a fornecer resultados heterogêneos e descontextualizados

[27] Busca feita em maio de 2007
[28] Termo escolhido aleatoriamente

quanto ao domínio de conhecimento. Nesses casos, o refinamento dos resultados dependerá da habilidade de os usuários transformarem suas questões em termos que delimitem o contexto no qual se espera informações.

Esse problema não se restringe aos sistemas genéricos de busca. Em consulta à Biblioteca Digital de Teses e Dissertações (BDTD) do IBICT, vê-se que uma busca pelo termo “informação” no campo ASSUNTO, retornou, nos primeiros resultados, teses ou dissertações sobre: indústria fonográfica, ecologia, interfaces de websites, marketing e sociedade da informação. Caberá ao usuário inserir novos termos que delimitem o domínio de conhecimento desejado. Ressalta-se que os dados da BDTD do IBICT, diferentemente dos dados da **PL**, são alimentados pelas instituições cooperantes (bibliotecas de universidade e instituições de pesquisa). No entanto, sabe-se que esses dados não são normalizados na BDTD. Infere-se daí que haverá grande revocação nos resultados de busca feitos a essa base. Seria recomendável que se utilizassem mecanismos de controle para obter resultados mais consistentes na exploração de uma base tão importante para a ICT brasileira.

O crescimento e o volume atual da ICT em meio eletrônico (mundial ou brasileira) são grandes o suficiente para justificar mudanças nas formas de tratamento e organização da informação. Contudo, é igualmente urgente refletir a respeito dos SICTs atuais, principalmente aqueles pautados na ausência de controle na alimentação de dados no sistema. O investimento que será demandado futuramente, para solucionar as inconsistências, talvez seja maior que os investimentos necessários para desenvolver sistemas previamente planejados para proporcionar maior confiabilidade no que diz respeito à organização da informação.

Na próxima seção serão apresentadas novas formas de organizar a informação em meio eletrônico, restringindo-se às que dizem respeito mais a este estudo sobre a **PL**.

3.4.1 Ontologias

Um importante fundamento relacionado à organização da informação na **PL** diz respeito às ontologias do sistema. Segundo a definição que consta na página do Conscientias[29] (CONSCIENTIAS, 2006):

> Uma ontologia caracteriza um acordo, o qual não necessariamente precisa abranger toda a conceituação de um determinado domínio, mas pode abranger apenas uma parte dele; ou seja, pode oferecer uma visão para o domínio. Dessa forma, uma ontologia atua como um contrato entre parceiros, permitindo que se comuniquem com segurança dentro do contexto do domínio de informação. Por exemplo, um agente de software que esteja comprometido com uma ontologia será capaz de interpretar semanticamente os itens de informação compreendidos por essa ontologia e se comunicar com outros agentes comprometidos com essa ontologia. Assim, uma ontologia estabelece uma comunidade de usuários na Internet.

A definição apresentada é restrita e aplicável a um domínio específico da área da Informática - essa percepção das ontologias é perceptível na **PL**. Porém o termo Ontologia não é novo, foi criado antes mesmo do desenvolvimento das tecnologias de informação. Segundo Lima-Marques (2006, p.17), etimologicamente o termo ontologia significa Ciência ou estudo do "ser" ou "ente", assim, a ontologia "é o estudo da existência de todos os tipos de entidades, abstratas ou concretas, que constituem o mundo".

Historicamente a ontologia tem origem na Grécia antiga com os pré-socráticos, mas Almeida e Bax (2003) explicam que para a organização da informação, o termo ontologia se diferencia daquele tradicional adotado na filosofia. Para esses autores, nas ontologias definem-se categorias para as coisas que existem em um mesmo domínio. Outra autora, Rios (2005, p.3), sistematizou conceitos de ontologia em diferentes domínios do conhecimento (Figura 8).

[29] A Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior (CONSCIENTIAS) foi criada para desenvolver ontologias que se prestem ao intercâmbio de informações entre agências de fomento e instituições ligadas ao tema Ciência, Tecnologia, Inovação e Informações de Aprimoramento de Nível Superior. Caracterizam-se como responsabilidades da Comunidade CONSCIENTIAS a concepção, elaboração, recomendação e manutenção das gramáticas relacionadas às ontologias submetidas pelas agências ou instituições conselheiras.

Área	Conceito
Filosofia	É a ciência que trata de seres em geral, enquanto seres, seus nomes e propriedades, concebidos como tendo uma natureza comum que é inerente a todos e a cada um dos seres.
Linguagem e cognição	Refere-se a tudo que existe no mundo composto por objetos, mudanças e relações entre eles. Ontologia pode ser baseada no mundo, na mente, no intelecto, na cultura ou na linguagem.
Medicina	È uma doutrina que estuda o ser da doença, como se a enfermidade existisse em conformidade a um tipo bem definido, a uma essência.
Sistemas de Informação	Segundo Chandrasekaran, Josephson e Benjamins, (1998)[30], ontologias são teorias de conteúdo sobre os tipos de objetos, propriedades de objetos e relacionamentos entre objetos que são possíveis em um domínio de conhecimento específico.
Inteligência Artificial	Guarino (1997)[31] define a ontologia como uma caracterização axiomática do significado do vocabulário lógico, e, para Sowa e Dietz (1999)[32], a ontologia define os tipos de coisas que existem no domínio de uma aplicação.

Figura 8 - Conceitos de ontologia em diferentes domínios do conhecimento (RIOS, 2005, p.3)

Um autor muito citado nas discussões sobre ontologias é Guarino (GUARINO, 1997), que percebe que há diferenças terminológicas consideráveis a respeito da ontologia e explica que o cerne da questão está na noção do que seja conceito. Na percepção de Alvarenga (2006, p.92), Guarino enfoca as ontologias sob o ponto de vista de sistemas baseados em conhecimento e discute como os princípios da ontologia formal podem ser usados na prática da engenharia do conhecimento.

De acordo com Campos (2004, p.25) a ontologia formal "é um formalismo classificado no nível ontológico, pois sistematiza conhecimento pretendendo a formalização de definições axiomáticas". Para um melhor entendimento da ontologia formal é importante antes entender um pouco dos mecanismos de representação do conhecimento. Campos (2004) entende que, no âmbito da Ciência da Computação, os referidos mecanismos auxiliam na implantação de estruturas computacionais. No âmbito da Ciência da Informação, contribuem para a elaboração de linguagens documentárias voltadas à recuperação e organização da informação. No âmbito da terminologia permitem a sistematização dos conceitos e elaboração de definições consistentes.

---

[30] CHANDRASEKARAN, B.; JOSEPHSON, J.; BENJAMINS, V. Ontology of Tasks and Methods. In: KAW,11.,1998, Alberta. **Workshop on Knowledge Acquisition, Modeling and Management**. Alberta. Banff, 1998.

[31] GUARINO, N. **Understanding, Building, and Using Ontologies**. International Journal of Human Computer Studies, Duluth, v. 46, n. 2-3, p. 293-310, fev./mar. 1997.

[32] SOWA, J., DIETZ, D. **Knowledge Representation: logical, philosophical, and computational foundations**. [s.l.]: Brooks Cole, 1999.

A representação do conhecimento pode ser classificada em quatro níveis: lógico, epistemológico, ontológico e conceitual. A seguir (Figura 9) é apresentado um quadro detalhando tais níveis.

NÍVEIS	CARACTERÍSTICAS	PRIMITIVAS	EXEMPLO
LÓGICO	É o nível da formalização, não há preocupação com a semântica em termos dos conceitos e de suas relações. O foco está em uma dada "sintaxe" que possibilite uma ação do pensar.	Predicados, funções.	$\forall x$ aluno(x) $\Rightarrow$ Corpo-acadêmico(x) $\exists x$ aluno(x) $\wedge$ Inteligente(x)
EPISTEMOLÓGICO	Neste nível a noção genérica de um conceito é introduzida como uma primitiva de estruturação de conhecimento.	Relações de estruturação	Aluno é uma subclasse do corpo acadêmico. Existem alunos que são inteligentes.
ONTOLÓGICO	Busca restringir o número de possibilidades de interpretação do conceito dentro de um dado contexto a partir de um formalismo que pretende representar o conteúdo do conceito.	Relações ontológicas	Todo aluno é um **objeto material.** Inteligente é uma **qualidade**.
CONCEITUAL	Independentemente de um formalismo, todo conceito possui uma interpretação definida. A estrutura dos conceitos em um determinado domínio está definida e o conhecimento é expresso na forma de uma especificação desta estrutura.	Relações conceituais	Nos exemplos a estrutura refere-se a interpretação de aluno num domínio acadêmico.

Figura 9 – Níveis da representação do conhecimento baseado em CAMPOS (2004, p.24-5) e Moreira, Alvarenga e Oliveira (2004).

Os níveis epistemológico e ontológico permitem a representação de conhecimento estruturado e formalizado.

> No nível epistemológico, especificam-se a estrutura dos conceitos e seus inter-relacionamentos. No nível ontológico, avança-se um pouco mais no processo de organização e classificação de um determinado domínio, e acrescenta-se a definição dos conceitos que nele estão inseridos. Enquanto o nível epistemológico é o nível de estruturação, o nível ontológico é o nível de significação (CAMPOS, 2004, p.25).

Para Campos (2004) a Ciência da Computação utiliza modelos de objetos e de dados para representações no nível epistemológico, contudo esses modelos são limitados para representar conhecimento. Por conta dessa limitação é que se introduziu no âmbito da Computação a noção de um nível ontológico. Na Ciência da Informação a teoria da classificação estaria, na visão de Campos (2004, p.25), em um nível de transição entre o nível epistemológico e ontológico: "apesar de não pretender chegar à definição dos conceitos de um dado domínio, ela possui um formalismo que possibilita a representação do conhecimento". Já as teorias do conceito e da terminologia podem ser classificadas como de um nível ontológico propriamente, pois permitem a sistematização de conhecimentos e possuem diretrizes para a elaboração de definições.

Outro ponto relevante para esta pesquisa sobre a **PL** é compreender as interpretações do termo Ontologia no conjunto de discussões que abordam aspectos da informação. Para tanto, foi usado o artigo de Moreira, Alvarenga e Oliveira (2004) que categorizou as interpretações em quatro grupos que serão a seguir detalhados:

a) Ontologia como um sistema conceitual subjacente a uma base de conhecimento[33].

b) Ontologia como um tipo especial de base de conhecimento.

c) Ontologia como um vocabulário usado por uma teoria lógica.

d) Ontologia como uma especificação de uma conceitualização.

A última interpretação da ontologia é a que mais se identifica com a **PL**, no entanto, todas serão detalhadas.

a) Ontologia como um sistema conceitual subjacente a uma base de conhecimento

Alguns pesquisadores não consideram as ontologias como objetos concretos, porém como uma estrutura conceitual subjacente a uma base de conhecimento. A ontologia precede a criação da base de conhecimento, é o conjunto de conceitos e relações que serão representados na referida base, e tal conjunto refere-se às ligações ontológicas desejadas. Uma vez que a ontologia pertence ao nível conceitual e não é apresentada de forma explícita no nível sintático, é possível que as sentenças em uma base de conhecimento estejam sujeitas a diferentes interpretações. Esta interpretação da ontologia como um sistema conceitual subjacente, por não estar situada no nível simbólico, não pode ser armazenada e operada computacionalmente.

b) Ontologia como um tipo especial de base de conhecimento

Alguns pesquisadores entendem a ontologia como uma base de conhecimento que se distingue das demais por: possuir apenas um tipo determinado de conhecimento; ou ser orientada a determinado tipo de tarefa. Em ambos os casos uma ontologia é um artefato concreto no nível simbólico e, portanto, pode ser compartilhada e transmitida. Ressalta-se apenas que no primeiro caso, a maioria dos autores entende que o conhecimento registrado

[33] As autoras usaram o termo "base de conhecimento" no sentido de um conjunto de sentenças descrevendo o estado de um domínio na forma de uma teoria lógica.

em uma ontologia deve descrever objetos e relações que estejam sempre presentes no domínio.

c) Ontologia como um vocabulário usado por uma teoria lógica

O estudo de Moreira, Alvarenga e Oliveira (2004) identificou que

> alguns pesquisadores classificam a ontologia como um artefato sintático, mas não exigem que ela tenha o rigor de uma teoria formal, enquanto outros definem ontologia como sendo apenas o vocabulário adotado em um domínio específico.

Em um sentido uma ontologia é um vocabulário de representação, frequentemente especializado para algum domínio ou assunto, em outro sentido uma ontologia é usada para referir a um corpo de conhecimento descrevendo algum domínio, tipicamente um conhecimento comum de um domínio, usando um vocabulário de representação.

Quando são exigidas definições formais dos termos e de suas relações, a interpretação corrente (da ontologia como um vocabulário usado por uma teoria lógica) coincide com a próxima interpretação (item d) abaixo), uma vez que os termos denotam conceitos e o registro das definições e relações, na forma de uma teoria formal, pode ser vista como uma especificação de uma conceitualização. Quando não se exige a representação das definições e das relações como uma teoria formal, a interpretação corrente possibilita visualizar a ontologia como um tipo de tesauro.

d) Ontologia como uma especificação de uma conceitualização

Esta interpretação, dentro da comunidade de representação de conhecimento, é a mais popular. Segundo Moreira, Alvarenga e Oliveira (2004), a definição mais famosa de ontologia diz o seguinte: “ontologia é especificação formal e explícita de uma conceitualização compartilhada" (Grubber citado por Moreira, Alvarenga e Oliveira, 2004). A figura (Figura 10) abaixo explora melhor os conceitos desta definição:

Formal	A ontologia pode ser expressa em uma linguagem formal.
Explícita	É um objeto de nível simbólico.
Compartilhada	O conhecimento é aceito por uma comunidade.
Conceitualização	Uma conceitualização é uma visão abstrata e simplificada do mundo que nós desejamos representar para algum propósito. Toda base de conhecimento, sistema baseado em conhecimento, ou agente atuando no nível do conhecimento é comprometido com alguma conceitualização, explicita ou implicitamente.

Figura 10 – Conceitos pertinentes a definição de ontologias de Grubber (citado por Moreira, Alvarenga e Oliveira, 2004).

Para ilustrar as noções de conceitualização serão usados exemplos a partir de uma visão formada pelo domínio da **PL** (Figura 11). Uma conceitualização desta visão poderia conter conceitos como: "título", "autor", "artigo", "compõe", "publica", “Produção bibliográfica”, “Currículo” etc. Uma especificação explícita desta conceitualização em lógica de primeira ordem poderia ser algo como:

Sentença em lógica	Linguagem natural
$\forall$x artigo(x) $\Rightarrow\exists$y (autor(y) $\wedge$ publica (x,y))	Para todo artigo existe um autor que o publica.
$\forall$x artigo(x) $\Rightarrow\exists$y (Produção bibliográfica (y) $\wedge$ parte-de(x, y))	Todo artigo é parte de uma produção bibliográfica.
$\forall$x título(x) $\Rightarrow\exists$y (Produção bibliográfica (y) $\wedge$ cursa(x, y))	Todo título compõe uma produção bibliográfica.
$\forall$x título(x) $\Rightarrow$ Currículo(x)	Todo título é um elemento do Currículo.
$\forall$x autor(x) $\Rightarrow$ Currículo(x)	Todo autor é elemento do currículo.

Figura 11 - Especificação Explícita de uma conceitualização

Diante da diversidade de compreensões sobre as ontologias, adotamos uma definição que entendemos ser a mais adequada para esta pesquisa sobre a **PL**. Desta forma, entende-se que as ontologias formalizam consensualmente a estrutura de conceitos dentro de um determinado domínio, e a partir desse consenso, se estabelecem regras para que as entidades que compõem esse domínio se relacionem. Com isso, sistemas de informação que adotam ontologias comuns podem compartilhar informações, e esse compartilhamento ocorre não apenas quando há equivalência entre termos idênticos, pois fragmentos do texto (que tanto podem ser uma frase, um parágrafo, ou uma seção), serão “entendidos” graças à semântica pré-estabelecida na estrutura do documento.

Comumente, as ontologias são organizadas no modelo de entidade-relacionamento. Esse modelo é constituído em classes (na terminologia computacional chamada também de conceitos) com definições de seus atributos e com os objetos que possuem estes atributos e integram estas classes. Entre as classes são criadas relações que possam existir entre os conceitos que ocorrem segundo um domínio particular de conhecimento, ou em alguma atividade específica. Assim, baseando-se em Almeida e Bax (2003) e em Pinto, Pereira e Burnham (2005) expõem-se alguns conceitos importantes das ontologias:

- Conceito: É algo que se deseja representar sobre determinado domínio. Como há diversos domínios é necessário delimitar um universo semântico, para tanto são arbitradas classes e categorias bem como as relações existentes entre elas.
- Classes: Descrevem conceitos de um domínio. Numa relação hierárquica, as classes podem se subdividir em níveis subordinados, e o principio (de subdivisão) que rege a distribuição dos elementos de uma classe é enunciado pela categoria. Assim, uma categoria mais específica herda as propriedades de uma categoria mais genérica até alcançar o nível da classe. Desta forma, na **PL**, a categoria de artigo publicado é sempre atribuída à classe produção bibliográfica.
- Atributos: São as características que descrevem os conceitos, ou seja, são as propriedades das classes e categorias. Ex: a categoria artigos completos possui 2 níveis específicos (dados básicos do artigo e detalhamento do artigo) e um dos atributos de dados básicos é titulo do artigo.
- Instância: São os conceitos e relações estabelecidos em uma ontologia específica. Assim, uma instância (dentro de um domínio) é um conceito que pertence a uma classe e que possui atributos específicos, segundo o direcionamento da instância. Exemplo de uma instância na **PL**: Produção bibliográfica de artigos publicados em 2006.
  - Produção: Bibliográfica
  - Tipo: Artigos Publicados
  - Periódico: Ciência da Informação
  - Ano: 2006

Estudos na área da Ciência da Informação condizentes às ontologias têm-se demonstrado atentos às diferenças conceituais em relação à Ciência da Computação. O trabalho de Moreira (2003) buscou relacionar dois instrumentos (os tesauros e as ontologias) usados na organização da informação, analisando definições sobre os dois instrumentos em estudos da área da Ciência da Informação e da Ciência da Computação. Identificou-se que há diferenças de propósitos entre os dois. O propósito dos tesauros é servir como instrumento de registro terminológico para ser usado por pessoas. Já as

definições sobre ontologia demonstram a necessidade de registro do conhecimento do domínio em uma linguagem que possa ser processada pelo computador para realizar inferências computacionais.

Segundo Moreira, uma ontologia é vista pela Ciência da Computação como um sistema de conceitos, da mesma forma que os tesauros. A diferença em relação aos tesauros pode ocorrer em termos de linguagem, de nível de formalização e de propósitos.

Por seu lado, Moreira, Alvarenga e Oliveira (2004) concluem que os tesauros da Ciência da Informação e as ontologias da Ciência da Computação possuem origens e propósitos distintos. O primeiro nasceu como um recurso auxiliar na indexação e busca de documentos; o segundo, para descrever os objetos digitais e suas relações. O que há em comum nessas origens é o fato de estarem relacionadas com a descrição de alguma entidade: assunto de uma área no primeiro caso e objetos e relações no segundo. Quanto às diferenças, aparentemente a Ciência da Computação entende que pode ser considerado ontologia tudo que modela um segmento da realidade. Por esta razão, em alguns textos da Ciência da Computação é comum se enquadrar os tesauros como ontologias terminológicas.

Alguns pesquisadores alegam que a distinção entre os tesauros e as ontologias da Ciência da Computação reside no fato das ontologias permitirem uma maior variedade de relações. Moreira, Alvarenga e Oliveira (2004) discordam explicando que tal visão

> advém da falta de entendimento do que é um termo e o que é relação segundo a teoria dos tesauros. Os tesauros, assim como algumas linguagens para representação de ontologias, apresentam um conjunto de relações pré-definidas para serem usadas para a estruturação dos conceitos. Este conjunto de relações de estruturação varia de tesauro para tesauro, em função da teoria subjacente e dos propósitos almejados. Já as relações observadas no domínio são representadas nos tesauros da mesma forma que qualquer outro conceito, enquanto que nas ontologias da Ciência da Computação, as relações são representadas de forma distinta das propriedades (isto é classes) e a elas podem ser atribuídas restrições e propriedades estruturais (e.g. transitividade) que podem ser usadas na realização de inferências.

O objetivo maior para a criação de ontologias, na opinião de Souza e Alvarenga (2004), parte da necessidade da existência de um vocabulário compartilhado para troca de informações entre comunidades, que podem ser formadas tanto por humanos como por agentes inteligentes. Já existem

ontologias ou projetos relacionados a elas; cita-se, a seguir, algumas consideradas interessantes:

- **DAML (http://www.daml.org/ontologies**) - Lista de ontologias disponíveis no site da DARPA Agent Markup Language (grupo composto por organizações interessadas em desenvolver tecnologias para a WEB). Entre as ontologias destaca-se a Unified Medical Language System (UMLS) criada por brasileiros;
- **WEBKB (http://www.webkb.org)** - Desenvolvida na Universidade de Griffith (Austrália), utiliza uma linguagem de representação de conhecimento que define associações e especializações entre termos predefinidos em uma única e ampla ontologia, projetada para facilitar a criação de outras ontologias;
- **OMV (http://omv.ontoware.org)** - Ontology Metadata Vocabulary, projeto que propõe um padrão de metadados por eles denominado como Vocabulário de Ontologia de Metadados. Apresenta ontologias detalhadamente descritas.
- **SWRC (http://ontoware.org/projects/swrc)** - Semantic Web for Research Communities, é uma ontologia que visa a modelar entidades de comunidades de pesquisa, incluindo publicações (metadados bibliográficos);
- **Open Biomedical Ontologies (http://obo.sourceforge.net)** - diversos vocabulários controlados, bem-estruturados para o uso compartilhado através de diferentes domínios biológicos e médicos;
- **Gene Ontology Home (http://www.geneontology.org)** – Dedicado à ontologia sobre genética. Possui o sistema de busca "Amigo" que possibilita visualizar a estrutura hierárquica do termo textualmente ou graficamente;
- **National Center for Biomedical Ontology (http://bioontology.org)** – Site bem estruturado voltado à organização da informação na área da biomédica. dedica-se essencialmente a estoques de informação produzidos a partir de pesquisas;
- **NLM/MESH (http://www.nlm.nih.gov/mesh)** - National Library of Medicine/ Medical Subject Headings - Não há uma ontologia, contudo o MESH está disponível em XML, o que facilita a formalização em ontologias na área médica para outras pessoas ou organizações que se dispuserem a criá-las;
- **MMI (http://marinemetadata.org)** – Marine Metadata Interoperability - Projeto de interoperabilidade de informações sobre estudos marinhos, possui um interessante tutorial para a elaboração de ontologias.

Igualmente importante é a Web Ontology Language (OWL), uma linguagem de marcação para publicação e compartilhamento de ontologias definidas pelo Web Ontology Working Group, que é parte do projeto da Web

Semântica da W3C[34]. Segundo Rios (2005), essa linguagem pode ser utilizada por aplicações que precisam não somente disponibilizar conteúdos, mas também processá-los. Na página da OWL[35] é chamada a atenção para o fato de que os recursos em OWL são direcionados ao desenvolvimento de ferramentas e ontologias para uso em comunidades específicas (particularmente nas ciências e no comércio eletrônico). Desta forma, tais recursos não têm como propósito serem compatíveis com uma arquitetura geral da WWW, porém com um conjunto mais restrito da Web Semântica.

Há uma perspectiva de que a convergência entre ontologias, linguagens de marcação e outras tecnologias facilite o desenvolvimento da Web Semântica, que segundo Berners-Lee (2001)

> [...] não é uma Web a parte, e sim uma extensão da atual, na qual a informação tem um significado bem definido, e tornará melhor a interação entre os computadores e as pessoas. Os primeiros passos para tecer a Web Semântica, dentro da estrutura existente da Web, já foram dados.

É provável que no âmbito da ICT a tendência seja o crescimento dessas relações "semânticas" entre os sistemas. Atualmente, a **PL** oferece recursos de interação (também chamados de interoperabilidade) com outros sistemas como a Biblioteca Científica Eletrônica Online (SciELO), como a base de patentes do Instituto Nacional de Propriedade Intelectual (INPI), o Diretório do Grupo de Pesquisas do CNPq, e alguns bancos de dissertações e teses de universidades.

Uma definição simplificada de interoperabilidade é vista em Marcondes e Sayão (2002, p.27) que falam na

> [...] possibilidade de um usuário realizar buscas a recursos informacionais heterogêneos, armazenados em diferentes servidores na rede, utilizando-se de uma interface única sem tomar conhecimento de onde nem como estes recursos estão armazenados.

Além dessa definição, esses autores destacam duas modalidades de interoperabilidade: uma com buscas distribuídas a diferentes servidores e outra com uma base de metadados centralizada. Na primeira, através da interface de busca, o usuário estipula o(s) termo(s) que o interessam e, após enviar os dados, o sistema se incumbe de distribuir a consulta a diferentes sites, segundo um protocolo padrão e os resultados são unificados e apresentados

---

[34] Consórcio formado por instituições acadêmicas, cientistas, empresas, profissionais e que estabelece padrões tecnológicos que regulam a WWW.

[35] http://www.w3.org/2004/OWL/

na tela com a formatação estipulada pelo sistema. Um dos protocolos utilizados nesse processo é o protocolo Z39.50, conhecido por proporcionar interoperabilidade entre catálogos automatizados de bibliotecas.

Na segunda alternativa há uma coleta periódica de metadados que são extraídos de documentos eletrônicos. Os metadados de diversos provedores de informação são compatibilizados através de protocolos padronizados e são coletados (harvesting) e armazenados em uma base centralizada de metadados (data warehousing), na qual são efetuadas as buscas de forma integrada.

3.4.2 As linguagens de marcação

O termo linguagem, no contexto das linguagens de marcação (do inglês *markup languages*), não tem relação com a linguagem debatida na área da lingüística. Aqui, as linguagens de marcação se inserem no âmbito da Informática e representam, de acordo com Bax (2001), um novo paradigma de gerenciamento (organização, recuperação e uso) da informação. Não há razões para denominar como novo paradigma de gerenciamento o que concretamente é um novo conceito computacional para estruturação de dados.

Uma evidência de que não se trata de um novo paradigma é o fato de que a linguagem XML (que será detalhada adiante) possui semelhanças com a norma ISO 2709. A ISO 2709 (Document Format for Bibliographic Interchange on Magnetic Tape) foi publicada em 1973 e atualizada em 1992. É voltada ao intercâmbio de informações bibliográficas em formato legível por computador. Foi criada a partir da necessidade de estabelecer padrões entre sistemas de bibliotecas para que os mesmos pudessem trocar dados através de arquivos seqüenciais, geralmente fitas magnéticas.

Esta norma não especificou o conteúdo nem o tamanho dos registros individuais, tampouco atribuiu significado específico aos designadores de conteúdo (tags em inglês). Entretanto, já estabelecia os conceitos de: registro, campos, características associadas aos campos, ordem dos campos e a idéia de tags para identificação dos campos, de forma semelhante à linguagem XML.

Os primeiros computadores, há 40 ou 50 anos, ofereciam um baixo nível de interação com as pessoas. Essa interação avançou bastante, alcançando o que a Informática denomina de um ‘alto nível de abstração’ que implica na

possibilidade de armazenar, organizar, recuperar e intercambiar informações. As linguagens de marcação contribuíram bastante para que isso ocorresse. Elas, segundo Bax (2001, p.32)

> permitem a construção de padrões públicos e abertos que estão sendo criados para se tentarem maiores avanços no tratamento da informação; elas minimizam o problema de transferência de um formato de representação para outro e liberam a informação das tecnologias de informação proprietárias.

Essas linguagens identificam, descritivamente, partes de um documento eletrônico, como: parágrafos, títulos, tabelas ou gráficos. A partir das descrições ou marcações dessas partes é possível fazer com que o computador identifique e "compreenda" a que se refere cada fragmento de um documento eletrônico. Assim, um documento eletrônico marcado em várias partes permite a um determinado programa de computador "entender" essas partes, possibilitando processar os documentos eletrônicos não somente como um todo, mas também de forma separada.

Há dois tipos básicos de marcação:

- **Procedimental (ou de procedimento)** - indicam como um programa processador de texto deve dispor o texto na página. Geralmente são sistemas de formatação proprietário, ou seja, cada software editor ou compilador de textos possui seu próprio conjunto de códigos que valem apenas para aquele sistema, que deverá rodar em um determinado sistema operacional ou em uma máquina específica.
- **Marcação descritiva (ou declarativa)** - essas linguagens usam marcas (ou tags) para caracterizar partes do documento para que elas sejam processáveis. Considera-se uma marca num documento, tudo aquilo que não for o conteúdo propriamente dito do documento. Com isso, as marcas indicam qual a função de cada parte de um documento, e não como o mesmo precisa ser visualmente apresentado (MÉNDEZ RODRÍGUEZ, 2002).

De acordo com Bax (2001), um documento é constituído por três componentes distintos: conteúdo, estrutura e estilo (ou formatação). O conteúdo é a informação propriamente dita, a estrutura define como se dá a organização do conteúdo, ou do conhecimento inscrito e o estilo define o aspecto físico, visual.

Outro aspecto importante destacado por Bax (2001) é que a utilização de padrões de marcação internacionais abertos (SGML, HTML, XHTML, XML, etc.), permitem a criação de documentos independentes de um determinado software, hardware, ou sistema operacional. Esses ainda podem ser interpretados por programas dos mais diversos ambientes computacionais, bastando que exista uma aplicação no ambiente que reconheça o padrão usado na criação do documento.

> Como são padrões abertos, a informação não fica aprisionada, pode-se desenvolver conversores de um padrão para outro. A aplicação que deve tratar a informação é que se encarrega de interpretar as marcas e processá-las, para efeitos de estilo, ou outros processamentos (BAX, 2001, p.34).

A **PL** é toda estruturada em linguagens de marcação como o XML, permitindo que todos os registros contidos em sua base sejam interpretados por navegadores da Internet ou programas editores de texto, ou ainda quaisquer outros aplicativos que sejam capazes de processar as linguagens de marcação. Os currículos da **PL** são estruturados conforme a ontologia desenvolvida pela CONSCIENTIAS. No ano de 2000 formou-se a Comunidade Linguagem de Marcação da Plataforma Lattes (LMPL) que, posteriormente, passou a se chamar Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior (CONSCIENTIAS)[36].

Segundo Mendez Rodriguez (2002) as origens do SGML (Standard Generalized Markup Language) remontam aos trabalhos de Charles Goldfarb, Edward Mosher e Raymond Lorie em 1970, mas foi somente em 1986 que o SGML se constituiu como um padrão internacional (a ISO 8879) para a descrição das linguagens de marcação e para a análise formal de documentos.

O SGML permite que se definam linguagens de marcação de forma independente, facilitando o intercâmbio e a conservação de recursos eletrônicos estruturados, por isso é vista também como uma (meta) linguagem, ou seja, uma linguagem para descrever outras linguagens[37]. Baseia-se em marcações genéricas, que identificam nos documentos as suas partes lógicas e/ou elementos que o constituem. Essas marcações definem nos documentos a estrutura e elementos "semânticos", que podem ser descritos indiferentemente da forma que esses elementos possam ser exibidos.

---

[36] A CONSCIENTIAS foi criada para desenvolver ontologias que se prestem ao intercâmbio de informações entre agências de fomento e instituições ligadas ao tema Ciência, Tecnologia, Inovação e Informações de Aprimoramento de Nível Superior.

[37] Um exemplo de derivada do padrão SGML é a linguagem HTML.

Ressalta-se que o SGML não é um conjunto predeterminado de marcações, mas uma linguagem que permite que sejam definidos conjuntos de marcações conforme necessidades específicas; o conjunto de todas as marcações passíveis de serem utilizadas por qualquer linguagem derivada do SGML é chamado de Document Type Definition (DTD).

Cada DTD estipula as regras de verificação para validar um documento. Desta forma, o DTD define quais elementos constituem a estrutura do documento (na **PL** poderia especificar, por exemplo, o título de um artigo, nome do periódico, o volume, o ano de publicação etc.) e o relacionamento (inclusive hierárquico) que existe entre estes elementos. Uma vez especificado um tipo de DTD para um documento, esse registro de DTD poderá ser usado para validá-lo, verificando-se se o conteúdo está adequado às regras daquele DTD específico.

A **PL** foi, inicialmente, desenvolvida a partir dos DTD, porém com a homologação[38] da linguagem XML Schema[39] pelo Consórcio W3C, a CONSCIENTIAS elaborou uma nova regra utilizando essa linguagem. Abaixo (Figura 12) apresenta-se, como exemplo, um fragmento em XML da **PL**:

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" DATA-ATUALIZACAO="16052007" HORA-ATUALIZACAO="130954" xmlns:lattes="http://www.cnpq.br/2001/XSL/Lattes">
+ <DADOS-GERAIS>
- <PRODUCAO-BIBLIOGRAFICA>
+ <TRABALHOS-EM-EVENTOS>
- <ARTIGOS-PUBLICADOS>
- <ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="49">
<DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="Análise da revista Ciência da Informação disponibilizada na SciELO a partir do seu vocabulário controlado" ANO-DO-ARTIGO="2002" PAIS-DE-PUBLICACAO="Brasil" IDIOMA="Português" MEIO-DE-DIVULGACAO="IMPRESSO" HOME-PAGE-DO-TRABALHO="" FLAG-RELEVANCIA="SIM" />
<DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="Transinformação" ISSN="01033786" VOLUME="14" FASCICULO="2" SERIE="" PAGINA-INICIAL="133" PAGINA-FINAL="138" LOCAL-DE-PUBLICACAO="Campinas" />
<AUTORES NOME-COMPLETO-DO-AUTOR="Fabio Mascarenhas e Silva" NOME-PARA-CITACAO="SILVA, F. M. e" ORDEM-DE-AUTORIA="1" />
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Ciência da Informação" PALAVRA-CHAVE-2="SCIELO" PALAVRA-CHAVE-3="Publicação Eletrônica" PALAVRA-CHAVE-4="" PALAVRA-CHAVE-5="" PALAVRA-CHAVE-6="" />
- <AREAS-DO-CONHECIMENTO>
</AREAS-DO-CONHECIMENTO>
<SETORES-DE-ATIVIDADE SETOR-DE-ATIVIDADE-1="Informacao e Gestao C&T" SETOR-DE-ATIVIDADE-2="" SETOR-DE-ATIVIDADE-3=""
</CURRICULO-VITAE>
```

Figura 12 - Exemplo de fragmento em XML da Plataforma Lattes

[38] Padronização feita pelo W3C, um consórcio mundial que define as regras (técnicas) de funcionamento da WWW.

[39] Apesar de possuir a mesma função da DTD, especificar a sintaxe de um documento XML, especifica também os tipos de dados de cada elemento desse documento. Com o XML Schema é possível ainda reutilizar a definição de elementos de outros esquemas, criar tipos de dados personalizados, especificar o número mínimo e máximo de vezes que um elemento pode ocorrer, criar listas e grupo de atributos. (FERNEDA, 2003).

É interessante explicar o HyperText Markup Language (HTML), que na definição de Toutain (2006, p.18-19) é

> a língua franca para publicação de documentos na Web. É um formato não-proprietário baseado no SGML e pode ser criado e processado por uma grande variedade de Ferramentas. O HTML utiliza Tags, como <h1> e </h1>, para estruturar o texto em cabeçalhos, parágrafos, listas, links de hipertextos, etc.

Sem dúvida, o HTML é a linguagem de marcação que mais ajudou na popularização da WWW, porém é limitada, tendo sido criada com o propósito de somente apresentar conteúdos. Foi uma linguagem de marcação bastante explorada, mas como afirma Bax (2001, p.36):

> Agora que as tecnologias voltadas a WWW estão relativamente maduras, as empresas estão procurando formas de introduzir maior flexibilidade em seus documentos (como suas páginas Web), para aumentar o potencial de troca de informações, visando ao comércio eletrônico, por exemplo. Entra em cena um novo padrão, a linguagem XML.

A eXtensible Markup Language (XML) apresenta-se como uma intermediária entre o SGML e o HTML, pois é uma metalinguagem com uma sintaxe específica e um conjunto de regras bem definidas. O XML encontra-se entre a complexidade do SGML e a simplicidade do HTML ou, como disse Edwards, citado por Bax (2001, p.36): “o XML oferece 80% da facilidade do SGML em 20% da complexidade do SGML”. O XML pode ser baseado em esquemas. O esquema é uma definição da estrutura de uma classe de documentos XML, onde o próprio esquema pode estar escrito ou não em sintaxe XML.

Algumas vantagens do XML são defendidas por Méndez Rodriguéz (2002):

- implica uma arquitetura da informação mais aberta e extensível, não necessitando versões diferentes que possam funcionar em futuros navegadores;

- os dados são compostos por múltiplas aplicações e a flexibilidade permite agrupar desde páginas da WWW até bases de dados;

- por ser uma metalinguagem hierárquica, possibilita que os dados em múltiplos níveis se integrem em um mesmo arquivo. Conseguir uma relação hierárquica como essa nos modelos atuais de base de dados - chamados de relacionais - implicaria relações complexas entre tabelas;

- motores de busca adaptados a nova linguagem extensível desenvolverão respostas mais adequadas e precisas, já que a codificação do conteúdo da WWW em XML define melhor a estrutura da informação;

- com relação aos metadados, o W3C está trabalhando em busca de uma maior consistência, homogeneidade e amplitude dos identificadores descritivos e das descrições de documentos XML através de RDF.

O Resource Description Framework (RDF) é, na visão de Marcondes (2006) uma aplicação especial para descrever recursos na WWW, e, assim como o XML, também é um padrão homologado pelo W3C. Para esse autor, enquanto o XML é uma linguagem genérica que estrutura documentos eletrônicos, o RDF é próprio para criar metadados com a finalidade de localizar e identificar recursos, por isso o RDF usa o XML dentro de um esquema bem mais estruturado.

O RDF baseia-se na concepção de que um documento web possui propriedades (ex: autor de um artigo, título do artigo, periódico de publicação do artigo, ano de publicação) e que toda propriedade possui atributos (“SILVA, F.M.”, “Análise da revista Ciência da Informação disponibilizada na SciELO a partir do seu vocabulário controlado”, “Transinformação”, “2002”). O valor de uma propriedade pode ser outro recurso: nesse exemplo, o valor da propriedade *autor* poderia ser o endereço da **PL** do autor disponível na WWW.

A literatura ressalta a flexibilidade como uma das principais vantagens da linguagem XML. Essa flexibilidade se refere à facilidade proporcionada aos desenvolvedores de sistemas (como a **PL**) para reutilizar os dados de uma base em XML para fins diversos. Os currículos da **PL**, por estarem no formato XML, podem ser usados na tabulação de indicadores ou na geração de um padrão de currículo personalizado, ou ainda para relacionar os conteúdos dos currículos com outros sistemas, seja de documentos da própria **PL** ou de outros sistemas disponíveis na Web.

Em um sistema como a **PL,** se houver divergência sintática entre os valores atribuídos para as propriedades, a organização da informação e respectiva possibilidade de recuperação estão comprometidas, indiferentemente da estrutura em XML que tiver sido estabelecida. Assim, os fundamentos básicos adotados para organização da informação com fins de

recuperação não se tornam inválidos em sistemas desenvolvidos a partir de linguagens de marcação.

Desta forma, no exemplo utilizado há pouco, se o autor Silva, F.M. digitou no nome do periódico “Trasnifnormção” em lugar de “Transinformação”, a estrutura pré-estabelecida em **XML** de nada adiantará, pois é a partir da grafia das palavras que a busca será efetuada.

Em análise bibliométrica feita por Silva (2004), avaliou-se a produção científica docente de um determinado programa de pós-graduação. Para tanto, foram utilizados dados da **PL**, mais especificamente do campo PRODUÇÃO BIBLIOGRÁFICA/ARTIGOS PUBLICADOS, e foram encontradas dificuldades, pois

> a Plataforma Lattes possui algumas limitações referentes à padronização dos registros. Com um olhar mais criterioso, é possível identificar algumas falhas que surgem no momento de recuperar as informações desejadas. É essencial que o preenchimento dos campos seja feito de forma cuidadosa e, se possível, padronizada. [...] É necessário padronizar o nome que o pesquisador utiliza em suas publicações. [...] Observamos, também, que não existe padronização na utilização de palavras-chave (SILVA, 2004, p.81).

Assim, percebe-se que um aprimorado sistema como a **PL**, compromete parte dos seus recursos ao desconsiderar princípios elementares de um sistema de RI. Não é possível inferir se, na fase de planejamento, os desenvolvedores da **PL** tenham previsto ou não esse problema, contudo, o fato é que a organização da informação na **PL** é racional, econômica e resulta em inconsistências, conforme será detalhado na seção 4.

Conforme explicado anteriormente, a **PL** se baseia na ontologia estabelecida pelo grupo CONSCIENTIAS e segue uma estrutura de relações hierárquicas. Entende-se que as relações hierárquicas são aquelas que se definem entre noções subordinadas em um ou mais níveis. Tais relações, sob um ponto de vista documental, são criadas entre noções, por isso compõem um sistema denominado nocional.

Para a ISO 1087, citado por Cintra et al (2002, p.50), um sistema nocional é “um conjunto estruturado de noções que reflete as relações estabelecidas entre as noções que o compõem e no qual cada noção é determinada pela sua posição no sistema”. Ainda conforme a ISO 1087, noção é a unidade de pensamento constituído por propriedades comuns a uma classe de objetos.

Nas relações hierárquicas há termos superiores a outros (superordenação), através de níveis arbitrariamente constituídos: tais níveis são coerentes para um dado sistema nocional construído, mas podem não fazer sentido para outros. Para Cintra et al (2002), uma macro-hierarquia (e também as hierarquias subseqüentes) de sistemas de classificação, como a Classificação Decimal de Dewey (CDD), tem como base uma organização lógico-hierárquica, e depende dos princípios ou características de divisão adotadas a partir dos objetivos que são desejados. A CDD por exemplo, destina-se a um universo de conhecimento global, enquanto que os tesauros voltam-se a domínios restritos e especializados de conhecimento.

Os sistemas de classificação citados (CDD e tesauros) estabelecem relações nocionais necessárias à organização de uma área, constituindo uma LD. Na **PL**, a hierarquia prevista na sua ontologia define a estrutura de organização da informação que poderá ser compartilhada com sistemas de informação de agências de fomento, ou quaisquer outras instituições nacionais de C&T.

Interessa perceber que, simultaneamente às mudanças ocorridas nos recursos de ontologias de linguagens de marcação para a ICT nacional, continua o processo de alimentação livre da base de currículos da **PL**. É preocupante perceber que os estoques de ICT em meio eletrônico crescem e estão sendo organizados sem considerar fundamentos que buscam proporcionar mais confiabilidade aos SRI. A análise objetivando verificar se, de fato, essa abertura compromete a consistência dos dados da PL, conforme hipótese norteadora desta pesquisa, será detalhada na próxima seção.

4 ANÁLISES DA PLATAFORMA LATTES

Os objetivos da PL são essencialmente informar o currículo de pesquisadores, ou seja, o que os mesmos fazem e fizeram e fornecer subsídios para elaborar políticas públicas ou diagnósticos da C&T brasileira. Para que esses objetivos sejam efetivamente alcançados é necessário introduzir mecanismos de controle na etapa de inserção dos dados, no processamento desses dados e na forma de apresentação dos currículos. Para sugerir aprimoramentos no sistema, analisou-se a **PL** em duas etapas: a primeira, a partir da lógica dos arquivos pessoais e a segunda, observando os procedimentos de preenchimento do sistema. Em ambas foram feitas análises críticas, porém na segunda foram também intercaladas sugestões direcionadas à **PL,** que podem ser incorporadas a qualquer outro SICT. Por fim, são apresentadas discussões finais e sugestões mais sistemáticas que completam estas análises.

4.1 A PLATAFORMA LATTES E A LÓGICA DOS ARQUIVOS PESSOAIS

Uma das funcionalidades da **PL** é gerar currículos que serão tornados públicos. Esses currículos são documentos que, à moda da **PL**, organizam referências a documentos (alguns públicos e outros privados) do arquivo pessoal, ou institucional, dos cientistas. Desta forma, na **PL** o currículo é um documento que estrutura os documentos/atividades dos usuários cadastrados.

Estes currículos servem a um delimitado segmento de atuação social: qual seja o segmento dos atores da C&T brasileira. Cada parte da estrutura dos currículos descreve atuações ou produções em C&T e cada usuário cadastrado preenche, individualmente, as atividades que foram por ele desenvolvidas. No caso das atividades desenvolvidas com a participação de outros autores é possível citá-los, entretanto será necessário que cada participante citado descreva, em seus respectivos currículos, a atividade comum a todos eles.

Numa ótica arquivística, cada currículo propõe uma organização do arquivo pessoal do pesquisador, porém Santos (2005) notou que muitos cientistas não se dão conta que sua produção documental possa servir, no futuro, como objeto de estudo e que esta produção permitiria estudar

> a evolução das políticas de pesquisa e de ensino científicos, a evolução desta ou daquela disciplina ou ainda o papel deste ou daquele cientista no desenvolvimento da ciência (CHARMASSON citado por Santos 2005, p. 23).

Contudo, é possível o seguinte questionamento: o que é arquivo pessoal e o que é arquivo institucional?

Como muitas pesquisas são desenvolvidas em laboratórios e muitos laboratórios financiam as pesquisas, esses mesmos laboratórios se apossam da documentação produzida pelo cientista. A esse respeito, Welfelé, citada por Santos (2005, p. 27), afirma existir um 'elo perdido' entre as esferas institucional e pessoal: o laboratório. Mas Welfelé encerra a discussão esclarecendo que os arquivos pessoais de cientistas são aqueles acumulados pelos cientistas, e os arquivos do laboratório são aqueles relacionados à continuidade da pesquisa científica, caso o cientista não faça mais parte da pesquisa.

Deste modo, Welfelé considera como arquivos pessoais de cientistas: correspondências; cadernos e cadernetas de laboratório e de experiências; dossiês de trabalho; notas de trabalho e de leitura; dossiês de artigos e obras (os manuscritos); notas de cursos; documentos de caráter biográfico; dossiês de caráter administrativo.

O currículo gerado na **PL** é um documento pessoal do cientista, pois não se trata de um documento de pesquisa, e sim de uma descrição ordenada e sistemática sobre pesquisas e quaisquer outras atividades e eventos que dizem respeito à vida acadêmica/profissional do pesquisador. Porém, a natureza pública da **PL** implica na seguinte situação: por mais que um determinado currículo seja um documento do arquivo pessoal de um indivíduo, esse mesmo documento compõe um acervo eletrônico público maior, de responsabilidade do CNPq.

Na percepção arquivística, um arquivo pessoal idealmente organizado requer uma análise das atividades realizadas pela pessoa da qual se organizará os documentos, os quais serão organizados e agregados em função das atividades exercidas pela pessoa que os acumulou ao longo da vida.

Mas a lógica da **PL** não prioriza este mesmo princípio, pois distribui as atividades em função do que a mesma representa, cada uma entendida isoladamente, desconsiderando o contexto no qual foi realizada. Assim, uma

palestra cadastrada na **PL** é somente uma palestra, o que impossibilita contextualizá-la no ambiente de uma pesquisa em curso, ou em uma atividade de extensão que tenha gerado o convite para a palestra.

Um exemplo prático: um pesquisador desenvolveu, ao longo de dois anos, uma metodologia para uso de indicadores bibliométricos na formulação de políticas públicas em C&T. Seu trabalho resultou na publicação de dois artigos e um livro, convites para proferir palestras em um congresso e um simpósio, e ainda um convite para ser consultor num programa de capacitação para técnicos do Ministério da Ciência e Tecnologia.

As atividades do exemplo citado permitem o registro de pelo menos sete atividades do pesquisador: um processo, dois artigos publicados, um livro, um curso de curta duração (com material didático), e duas palestras. Todas são desdobramentos decorrentes da criação de uma metodologia específica. Ao preencher o currículo, as atividades são desmembradas de um núcleo de ação que originou um conjunto de ações, ou seja, são descontextualizadas. Os artigos serão então incluídos em um conjunto de artigos publicados e o mesmo ocorrerá com as demais atividades.

Vê-se que a menção a cada uma dessas atividades remove das mesmas seu significado, pois elas não foram geradas de forma descontextualizada, mas dentro de um contexto bem definido. E ainda, o registro de cada atividade, isoladamente, contribui para tornar o currículo demasiadamente longo e pouco informativo, em decorrência da falta de contextualização das atividades arroladas.

O quadro (Figura 13) abaixo representa como seriam distribuídas as atividades do pesquisador do exemplo citado:

CLASSES	SUBCLASSES	ATIVIDADES
TRABALHOS EM EVENTOS	**Artigos publicados em periódicos**	2 artigos publicados
	Livros e capítulos	1 livro publicado
PRODUÇÃO TÉCNICA	**Processos ou técnicas**	1 metodologia desenvolvida
	Demais tipos de produção técnica	
	Cursos de curta duração ministrados	1 curso ministrado
	Desenvolvimento de material didático ou instrucional	1 apostila para o curso
	Apresentações de trabalho	2 palestras

Figura 13 – Atividades do pesquisador do exemplo

O princípio de categorização da **PL** é baseado numa visão dualista, que entende as atividades como produção ou atuação. Essa divisão define as regras de relações entre as classes e hierarquias que compõem a **PL**. Numa visão arquivística, tais relações desfavorecem a constituição de arquivos pessoais, pois segrega em partes o que em vida se realizou de forma articulada ou contextualizada.

Essa divergência da **PL** com relação aos princípios arquivísticos compromete um dos objetivos da plataforma, que é apresentar em formato organizado e padronizado os currículos dos pesquisadores. Como conseqüência, o currículo de um pesquisador experiente e com uma produção representativa, sob o ponto de vista quantitativo, pode ser bastante extenso. Quanto mais informações no currículo, mais difícil será fazer uma leitura sistêmica dele. Por sistêmica, referimo-nos a uma análise conjuntural da vida do pesquisador. Desta forma, a PL, ao ignorar a lógica arquivística, prejudica a compreensão da atuação do pesquisador e, neste sentido, compromete o objetivo mencionado no início deste parágrafo.

Apenas para fins de demonstração, usou-se o currículo de um determinado pesquisador que foi acessado em 04/06/2007[40]. Os números expressam uma produção admirável: cento e noventa e três artigos, seis livros, e cinqüenta e oito trabalhos publicados, entre outros itens. Impresso, totaliza 17 páginas contendo apenas informações da atuação como docente e publicações.

Uma breve leitura do referido currículo revela a dificuldade para contextualizar os itens. A disposição das atividades em tópicos separados conduz ao entendimento de cada produção bibliográfica e atividade docente como atuações distintas e não inter-relacionadas. À medida que as páginas são roladas na tela do computador aumenta a sensação de ler uma simples listagem, e não ter acesso a um “espelho” da atuação do pesquisador. Um dos problemas está na seqüência cronológica das ações referenciadas: o último registro do tópico **publicação de artigos** tem data de 1967, e em seguida é apresentado no tópico **livros publicados** uma publicação do ano de 2005.

[40] A última atualização deste currículo ocorreu em maio de 2006.

A formatação[41] do currículo gerado a partir da base de currículos da **PL** não é rígida, logo, são possíveis variações na exibição das informações sem afetar o conteúdo dos currículos, através da seleção das informações a visualizar. Isso é possível devido ao fato de os currículos da **PL** estarem estruturados no padrão XML, razão pela qual os conteúdos podem ser retrabalhados em outras aplicações, flexibilizando a utilização para diversos fins.

Mas há limites na flexibilidade para utilização dos dados, sendo um deles a impossibilidade de contextualizar as partes que compõem o currículo. O "desenho" do sistema conduz os usuários a registrar suas atividades de forma descontextualizada, ou ainda, desencoraja-os a inserir de forma completa as ações por ele exercidas como um ator no contexto da C&T. Trata-se de um problema de planejamento, para o qual uma ação corretiva talvez não seja suficiente, sobretudo em razão do sistema não ter sido concebido para contextualizar as atividades dos atores.

Torna-se difícil, àqueles que preenchem o sistema, explicitar os vínculos que de fato existiram em suas vidas. O currículo do pesquisador usado como exemplo, mostra um docente, com doutorado no Massachusetts Institute Of Technology (MIT) há 30 anos, e com excelente produção científica. Mas, supostamente revela que ele não participou de eventos e/ou não exerceu quaisquer cargos de políticas públicas. Não se sabe se ele deixou de inserir tais informações ou se fato ele jamais participou de eventos e/ou ocupou cargos públicos.

Para os arquivos pessoais de cientistas é importante identificar as funções das atividades exercidas pelas pessoas durante suas vidas. Isso possibilita distinguir, por exemplo, a atuação de um determinado indivíduo como cientista ou como pessoa pública. Como pessoa pública, podem haver muitos registros que permitam contextualizar sua atuação. Agrupá-los e dar sentido a esses registros como um arquivo pessoal pode demandar um esforço que dependerá da dispersão dos documentos em diferentes instituições.

Com um acervo de mais de um milhão de currículos[42], a base da PL, apesar de por definição ser constituída por documentos biográficos de atores

[41] Disposição estética do documento.
[42] Em agosto de 2007 a **PL** ultrapassou um milhão de currículos.

da C&T, não se caracteriza como uma fonte ideal para a organização de documentos que retratam as atividades de pesquisadores. A forma como foi desenvolvida privilegia uma distribuição das atividades exercidas pelo sujeito de forma isolada e descontextualizada, tornando necessário registrar repetidas vezes um conjunto de ações que, originalmente, ocorreram de forma concatenada. Isso resolve um problema: o registro de todas as ações. Mas acarreta outros: a produção de um documento biográfico demasiadamente longo e pouco informativo.

4.2 ANÁLISE DO PREENCHIMENTO DA PLATAFORMA LATTES

A forma de cadastro e respectivo preenchimento dos campos da **PL** foram modificados ao longo dos anos e continua em processo de aperfeiçoamento. Dito isso, é importante entender que as interfaces de preenchimento disponíveis no período das análises[43] podem sofrer alterações com o passar do tempo.De todo modo, buscou-se convergir as avaliações críticas aos aspectos estritamente relacionados à organização da informação.

Atualmente,[44] a **PL** é dividida em sete módulos; são eles:

- **dados gerais:** concentra os dados de identificação, os endereços, a formação acadêmica e complementar, a atuação profissional, as áreas de atuação e os prêmios e títulos honoríficos;
- **produção bibliográfica:** concentra toda a produção bibliográfica realizada, artigos completos, livros, textos em periódicos, traduções, partituras, etc;
- **produção técnica:** concentra toda a produção técnica do usuário, softwares, produtos, trabalhos técnicos, maquetes, etc;
- **orientações:** módulo destinado a todas as orientações ou supervisões (concluídas ou em andamento);
- **produção cultural:** concentra toda atividade relacionada à área cultural, apresentações de obras, arranjos ou composições musicais, artes visuais, programas de rádio ou tv, etc;
- **eventos:** concentra informações relacionadas à participação em eventos como palestras, seminários, etc;

[43] Análises feitas no período de junho a agosto de 2007
[44] Junho de 2007

- **bancas:** concentra informações relacionadas à participação em bancas e comissões julgadoras.

Cada módulo contém diferentes campos que permitem ao usuário inserir conteúdos em forma de texto. Para fins desta pesquisa optou-se em categorizar as formas de preenchimento da **PL** em três grupos: **Autonomia Total**, **Autonomia Parcial**, e **Sem Autonomia**. Apesar de o sistema não ser assim subdividido, tal classificação é proposta para sistematizar as análises e discussões. Contudo, salienta-se que os campos são inter-relacionados e as inconsistências encontradas também, logo, tanto os problemas como as sugestões de melhoria poderão dizer respeito a mais de uma categoria. As características essenciais de cada categoria (e que serão desenvolvidas nas seções 4.2.1, 4.2.2 e 4.2.3, respectivamente) são:

a) Autonomia Total: O usuário tem a liberdade de cadastrar as palavras que desejar, sem restrição ou qualquer direcionamento.

Exemplos de campos com Autonomia Total: Título de uma publicação, Nome de autores, palavras-chave.

b) Autonomia Parcial: São campos que, inicialmente, têm autonomia total, porém, cada novo termo cadastrado pelo usuário é automaticamente armazenado no sistema, que vai criando uma lista de termos exclusiva (Figura 14) do usuário. Nas próximas vezes que o usuário inserir outros termos, será possível consultar e adotar termos anteriormente criados por ele. O usuário pode excluir qualquer termo dessa lista.

Exemplos de campos com Autonomia Parcial: palavras-chave, nome de autores.

Figura 14 - Lista de Termos

c) Sem Autonomia: O sistema prevê, inicialmente, opções que o usuário deve selecionar. Entretanto, a existência dessas opções prévias não impede que novos termos sejam incluídos, caso o usuário não se satisfaça com as opções oferecidas.

Exemplos de Campos Sem Autonomia de Preenchimento: idioma de publicação, Título de periódico/ISSN, Áreas do Conhecimento, Setores de aplicação.

Para a análise de cada categoria utilizaram-se currículos consultados na **PL**, que apesar de não terem sido coletados segundo procedimentos de amostragem, não foram aleatoriamente escolhidos, mas sustentados pela estratégia desenvolvida pela SciELO. Para aspectos da organização da informação postulou-se que as partes do currículo referentes à produção bibliográfica fossem mais apropriadas, pois exigem representações conceituais mais complexas se forem comparadas, por exemplo, a dados pessoais do pesquisador.

Pensando na possibilidade de usar outro sistema como parâmetro, para fins comparativos, optou-se por utilizar exemplos retirados de periódicos disponíveis na SciELO[45]. A escolha deu-se em razão da credibilidade alcançada por este sistema na comunidade científica brasileira (e internacional). Para alcançar a credibilidade atual, a SciELO precisou estipular

[45] Detalhes sobre a SciELO na Seção 2.2.

critérios[46] claros de admissão e manutenção dos periódicos em sua coleção. Para um periódico ser aceito na SciELO é necessário, no mínimo, que obedeça pré-requisitos como: ser um periódico de caráter científico; haver arbitragem por pares; possuir um conselho editorial; manter periodicidade mínima; ter publicado no mínimo quatro números; manter a pontualidade nos lançamentos de novos números; manter cadastro dos autores; seguir normalização; e, conter título, resumo e palavras-chave no idioma do texto do artigo e no idioma inglês, quando este não é o idioma do texto.

Destaca-se que o periódico que almejar sua inserção na SciELO deverá explicitar qual (quais) a(s) norma(s) seguida(s) para a apresentação e estruturação dos artigos, e também para elaboração das referências bibliográficas e das palavras-chave. Assim, evidencia-se a submissão dos periódicos disponibilizados na SciELO a padrões internacionais que servem a propósitos de tratamento da informação.

Para a seleção dos periódicos considerou-se a quantidade de fascículos já publicados, fator esse que evidencia a consolidação do periódico perante os pares. A escolha foi feita a partir da lista denominada COLEÇÃO DA BIBLIOTECA, que apresenta os periódicos disponibilizados na SciELO, ordenados em oito categorias[47]. Para cada categoria foi selecionado um periódico e, com isto, as categorias e respectivos periódicos assim foram arroladas:

- *Ciências Agrárias* (Arquivo Brasileiro de Medicina Veterinária e Zootecnia);
- *Ciências Biológicas* (Memórias do Instituto Oswaldo Cruz);
- *Ciências da Saúde* (Arquivos Brasileiros de Cardiologia);
- *Ciências Exatas e da Terra* (Brazilian Journal of Physics);
- *Ciências Sociais Aplicadas* (Ciência da Informação);
- *Engenharias* (Brazilian Journal of Chemical Engineering);
- *Linguistica, Letras e Artes* (DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada);
- *Humanas* (Estudos Avançados).

[46] Os critérios estão disponíveis no endereço http://www.scielo.br/criteria/scielo_brasil_pt.html.
[47] Esta categorização é da própria SciELO e serve basicamente para listar os periódicos segundo grandes áreas de conhecimento.

Para uniformizar o período dos artigos usados na análise definiu-se o primeiro número dos periódicos publicados no ano de 2006, ou seja, ao final foram analisados 8 números de periódicos, sendo o primeiro número de 2006 de cada um dos 8 selecionados. Considera-se que o espaço de tempo entre a publicação dos artigos e a coleta desta análise tenha sido suficiente para que os autores tenham cadastrado os referidos artigos publicados nos seus currículos da **PL**.

Ao todo, a análise foi elaborada a partir de 80 artigos, publicados por 282 autores. Há procedimentos específicos de análise para cada uma das três formas de preenchimento (Autonomia Total, Autonomia Parcial, Sem Autonomia), entretanto alguns processos foram comuns às três:

- Após definidos os periódicos e números da SciELO utilizados na análise, acessou-se cada artigo desses periódicos;
- em cada artigo foi (foram) copiado(s) o(s) nome (s) do(s) autor(es), para em seguida consultar, na página de busca da **PL**[48], o currículo do pesquisador;
- os currículos visitados (sem aparente razão) omitiam às vezes os dados completos. Em alguns casos não revelavam, por exemplo, as palavras-chave. Para solucionar tal problema acrescentou-se ao endereço do currículo do pesquisador o comando "**&tipo=completo"**;
- os autores que não tinham currículos (comumente os estrangeiros) e aqueles que não registraram os currículos pesquisados, obviamente não foram considerados.

4.2.1 ANÁLISE DOS CAMPOS COM AUTONOMIA TOTAL

Para a análise da categoria dos campos com Autonomia Total consideraram-se inconsistentes os dados preenchidos nos currículos de forma diferente da produção bibliográfica que foi registrada no periódico da SciELO. Para verificar este aspecto, considerou-se o campo título como o mais adequado, pois, sob o ponto de vista sintático, só pode haver equivalência entre dois registros de títulos quando ambos forem idênticos, diferentemente de uma avaliação semântica, na qual um mesmo significado pode estar presente em termos diferentes.

[48] http://buscatextual.cnpq.br/buscatextual/index.jsp

Os exemplos retirados dos periódicos analisados estão dispostos no formato do modelo abaixo:

MODELO
Exemplo: Nome do periódico.
Título Original do artigo: O título conforme consta no original disponível na SCIELO.
Autor: Forma como o autor cadastrou o artigo em seu currículo na **PL**.
Problemas: Inconsistências verificadas na comparação entre o título original e o cadastrado na **PL**.

Exemplo 1: Ciência da Informação
Título Original do artigo: Ciência da informação e cognição humana: uma abordagem do processamento da informação.
Autor: Cencia da linformação e cognição: uma abordagem do processamento da informação.
Problemas: Erro de digitação (CENCIA DA INFORMAÇÃO em lugar de CIÊNCIA DA INFORMAÇÃO); dados incompletos (falta a palavra HUMANA).

Exemplo 2: DELTA (Documentação de Estudos em Lingüística Teórica e Aplicada)
Título Original do artigo: Análise de conteúdo e análise do discurso: o lingüístico e seu entorno.
Autor 1: Análise de conteúdo e Análise do discurso: o lingüísticvo e seu entorno.
Autor 2: Análise de conteúdo e Análise do discurso: o lingüísticvo e seu entorno.
Problema: Erro de digitação (LINGÜÍSTICVO em lugar de LINGÜÍSTICO).

Exemplo 3: Arquivo Brasileiro de Medicina Veterinária e Zootecnia
Título Original do artigo: Coinfecção experimental de circovírus suíno tipo 2 isolado no Brasil e parvovírus suíno em suínos SPF.
Autor 1: Coinfecçãoexperimental de circovírus suíno tipo 2 isolado no Brasil e parvovírus suíno em suínos SPF.
Autor 2: Coinfecção experimental de circovírus suíno tipo 2 (PCV2) isolado no Brasil e parvovírus suíno (PPV) em suínos SPF.
Autor 3: Coinfecção experimental de circovírus suíno tipo 2 (PCV2) isolado no Brasil e parvovírus suíno (PPV) em suínos SPF.
Problemas
Autor 1: Erros de digitação (COINFECÇÃOEXPERIMENTAL em lugar de COINFECÇÃO EXPERIMENTAL)
Autor 2 e 3: Inclusão de palavras inexistentes no original ((PCV2) e (PPV)).

Exemplo 4: Arquivo Brasileiro de Medicina Veterinária e Zootecnia
Título Original do artigo: Mistura de proteínas morfogenéticas ósseas, hidroxiapatita, osso inorgânico e colágeno envolta por membrana de pericárdio no preenchimento de defeito ósseo segmentar em coelhos.
Autor 1: Mixture of bone morphogenetic protein, hydroxyapatite, inorganic bone and collagen interposed by pericardium barrier membrane in the filling of the segmental bone defect in rabbits.
Problema: Idioma no CV Lattes diferente do original do periódico.

Exemplo 5: Memórias do Instituto Oswaldo Cruz
Título Original do artigo: Identification of sex pheromones of Lutzomyia longipalpis (Lutz & Neiva, 1912) populations from the state of São Paulo, Brazil.
Autor 1: Identification of sex pheromones of Lutzomyia longipalpis (Lutz & Neiva) populations from the state of São Paulo.
Autor 2: Identification of sex pheromones of Lutzomyia longipalpis (Lutz & Neiva, 1912) populations from São Paulo State, Brazil.
Problema
Autor 1: Dados incompletos (não incluiu 1912 nem Brazil)
Autor 2: Erro de digitação no final do título (from São Paulo State, Brazil).

Exemplo 6: Memórias do Instituto Oswaldo Cruz
Título Original do artigo: Taeniosis-cysticercosis complex in individuals of a peasants' settlement (Teodoro Sampaio, Pontal of Paranapanema, SP, Brazil)
Autor 1: Taeniosis-cysticercosis complex in individuals of a peasants
Autor 2: Taeniosis-cysticercosis complex in individuals of a peasant´settlement (Teodoro Sampaio, SP, Brazil).
Autor 3: Taeniase-cysticercosis complex in individuals of pe.josimo peasants' settlement (Teodoro Sampaio, Pontal of Paranapanema-SP-Brazil)
Problemas: Dados incompletos, erros de digitação.
Autor 1: O título está incompleto (faltou settlement (Teodoro Sampaio, Pontal of Paranapanema, SP, Brazil))
Autor 2: Dados incompletos (PONTAL OF PARANAPANEMA).
Autor 3: Incluiu palavra inexistente (pe. josimo).

Exemplo 7: Brazilian Journal of Chemical Engineering
Título Original: The effects of sucrose on the mechanical properties of acid milk proteins-k-carrageenan gels.
Autor: Influence of sucrose on the mechanical properties of acid milk protein-k-carrageenan gels.
Problema: Substituição de palavra (INFLUENCE em lugar de EFFECTS).

Exemplo 8: Brazilian Journal of Chemical Engineering
Título Original: Application of interval analysis for gibbs and helmholtz free energy global minimization in phase stability analysis.
Autor: Application of Interval Analysis for Gibbs and Helmholtz FreeEnergy Global Minimization in Phase Stability Analysis.
Problema: Erros de digitação (FREEENERGY em lugar de FREE ENERGY).

Exemplo 9: Brazilian Journal of Physics
Título Original: Electron spin resonance dating of shells from the sambaqui (shell mound) Capelinha, São Paulo, Brazil.
Autor: Electron Spin Resonance dating of shells.
Problema: Dados incompletos (FROM THE SAMBAQUI (SHELL MOUND) CAPELINHA, SÃO PAULO, BRAZIL.)

Exemplo 10: Brazilian Journal of Physics
Título Original: Thermo-statistics of irreversible processes: a Boltzmann-Gibbs-style ensemble formalism.
Autor: THERMO-STATISTICS OF IRREVERSIBLE PROCESSES
Problema: Dados incompletos (A BOLTZMANN-GIBBS-STYLE ENSEMBLE FORMALISM.)

Exemplo 11: Arquivos Brasileiros de Cardiologia
Título Original: Respostas cardiopulmonares ao exercício em pacientes com insuficiência cardíaca congestiva de diferentes faixas etárias.
Autor 1: Rspostas cardiovasculares ao exercício em paciente com insuficiência cardíaca congestiva de diferentes faixas etárias.
Autor 2: Respostas Cardiovasculares ao Exercício em Pacientes com Insuficiência Cardíaca Congestiva de Diferentes Faixas Etárias.
Problemas
Autor 1: Erros de digitação (RSPOSTAS em lugar de RESPOSTAS); substituição de palavras (CARDIOVASCULARES em lugar de CARDIOPULMONARES)
Autor 2: substituição de palavras (CARDIOVASCULARES em lugar de CARDIOPULMONARES).

Exemplo 12: Arquivos Brasileiros de Cardiologia
Título Original: Estudo "LOTHAR": avaliação de eficácia e tolerabilidade da combinação fixa de anlodipino e losartana no tratamento da hipertensão arterial primária.
Autor 1: The LOTHAR study: evaluation of efficacy and tolerability of the fixed combination of amlodipine and losartan in the treatment of essential hypertension.
Problema: Idioma no CV Lattes diferente do original do periódico.

Para a **PL** (e para qualquer outro SICT), a diferenciação entre palavras pode ocasionar inconsistências e uma delas está relacionada à recuperação da informação. Os exemplos indicam situações nas quais haveria comprometimento nos resultados numa busca por determinados termos.

Um dos campos da página de busca da **PL** é o de ASSUNTO que, de acordo com explicações do próprio sistema, faz busca nos campos de título e das palavras-chave da produção científica, tecnológica e artística do pesquisador. Se fosse feita uma busca pelo termo LINGÜÍSTICO, certamente seriam recuperados diversos currículos, entretanto não é certo se os pesquisadores do exemplo 2 estariam incluídos no resultado[49], pois o artigo "Análise de conteúdo e análise do discurso: o lingüístico e seu entorno", foi cadastrado nos currículos de seus autores na seguinte forma: "Análise de conteúdo e Análise do discurso: o lingüísticvo e seu entorno". Essa diferença sintática provocada por erro de digitação impossibilita combinar o termo da busca com os registros existentes no sistema.

Nos exemplos 4 e 12 há uma situação diferente: o título original do artigo está em português, mas os autores o cadastraram em inglês na **PL**. Para fins de recuperação da informação, as palavras cadastradas em inglês apenas serão úteis para estratégias de buscas formuladas com termos na língua inglesa. É importante ressaltar que as buscas feitas na **PL** normalmente o são

[49] Acrescenta-se o fato de os autores também não terem cadastrado o termo LINGÜÍSTICO no campo das palavras-chave.

em língua portuguesa, ou seja, se o título original do artigo estiver em inglês (algo comum na literatura estrangeira e em alguns casos da brasileira também), haverá comprometimento nos resultados.

Nos exemplos 9 e 10 os autores omitiram nos seus currículos partes do título original do artigo. Com isso, uma busca contendo os termos não mencionados pelos autores trará prejuízos ao resultado. Salienta-se que na análise feita no periódico *Estudos Avançados* não houve divergência entre os títulos dos artigos no periódico daqueles cadastrados pelos autores na **PL**. Entretanto, o fato de apenas 7 dos 21 autores (soma de todos os artigos publicados no primeiro fascículo de Estudos Avançados de 2006) terem cadastrado os respectivos artigos na **PL** pode ter influenciado neste resultado.

Sobre os problemas apontados, é salutar citar, brevemente, aspectos da interoperabilidade (também chamado “enlaces”), que permitem um sistema compartilhar/usufruir recursos de outros sistemas. O compartilhamento ocorre graças a padrões de protocolo de comunicação e padrões de organização de dados. O primeiro remete a aspectos mais técnicos, o segundo aos conteúdos que são inseridos nos sistemas. Tornar um sistema “interoperável” com outro - em termos de protocolos de comunicação - assegura a troca de sinais entre duas ou mais máquinas. Mas, para haver um intercâmbio de conteúdos humanamente inteligíveis, é necessário também haver compatibilidades sintáticas/semânticas nos textos.

Se a interoperabilidade for efetuada através do campo título, haverá problemas. No exemplo 11 há uma situação em que o título do artigo no periódico difere dos cadastrados pelos dois autores. Percebe-se que, além do erro de digitação, houve também a troca do termo “CARDIOPULMONARES” por “CARDIOVASCULARES”. Há registro deste artigo em pelo menos três sistemas diferentes: a **PL**, a SciELO e a LILACS. Na SciELO e na LILACS os títulos estão corretos e idênticos, diferentemente de currículos da **PL**.

A **PL** é interoperável com a SciELO e com a LILACS, permitindo enlaçar um sistema a outro. Segundo Santana et al (2001), os enlaces são estabelecidos entre os textos na SciELO e os seus respectivos currículos por meio dos nomes de autores. Desta forma, um nome em um currículo indica a SciELO se o mesmo é um dos autores de artigos. E nos artigos da SciELO, os

autores que possuem currículos na **PL** têm seus nomes ligados aos respectivos currículos.

O procedimento ocorre da seguinte forma:

> a Bireme envia, periodicamente ao CNPq para processamento, um arquivo extraído da SciELO, com registros contendo, cada um deles, os autores (como são citados), o título e a URL do artigo. Neste processamento, para cada artigo e autor, procuram-se quais são os detentores de currículos Lattes cujos nomes são compatíveis com o nome de citação. Os currículos assim selecionados são examinados para descobrir, através de comparação não exata, em qual deles está mencionado o artigo que se está processando. Quando encontrado, acrescenta-se a URL do currículo selecionado ao registro enviado pela Bireme. (SANTANA et al, 2001, p.49).

A partir desse procedimento, são geradas duas tabelas, uma é enviada ao CNPq, que passa a ter uma lista de currículos de autores com artigos na SciELO. E a outra é enviada à BIREME que tomará conhecimento dos artigos da SciELO cujos autores possuem currículo cadastrados na **PL**.

Para o enlace com a LILACS, o procedimento adotado foi semelhante. Este, apesar de simples é, na visão de Santana et al (2001), oneroso do ponto de vista operacional. Na interoperabilidade com a base de dados do INPI e com o Diretório de Grupos de Pesquisa do próprio CNPq (que faz parte da **PL**) o enlace é mais simples, pois as ligações são feitas a partir do Cadastro de Pessoa Física (CPF) de cada pesquisador. Como a numeração do CPF é única para cada cidadão, as chances de inconsistências são minimizadas. Vale lembrar que o uso do CPF como elemento comum no enlace entre sistemas vale apenas para autores brasileiros.

Foi comum verificar que, tanto no periódico como no currículo, há autores que optam por registrar seus nomes próprios de formas diferentes. Percebeu-se também, que foi recorrente encontrar artigos na SciELO de autores que eram cadastrados na **PL** mas o enlace não foi criado na página do artigo do periódico. Para constatar esse fato foi necessário efetuar uma busca na página da **PL** com os nomes dos autores, e então verificar que faltava o enlace na SciELO. Um dos motivos desta falha se explica certamente pela diferença nos nomes dos autores. A seguir, apontam-se algumas diferenças identificadas no conjunto de currículos avaliados:

NOME DO AUTOR NO ARTIGO DO PERIÓDICO	NOME DO AUTOR NA PL
Décio Rocha	Décio Orlando Soares da Rocha
W. D. Marra Jr	Wiclef Dymurgo Marra Junior
J. Belincanta	Juliana Belincanta Ximenes
Jairo Pinheiro	Jairo Pinheiro da Silva
Rosângela Cipriano	Rosangela Cipriano de Souza
O. Baffa	Oswaldo Baffa Filho
Christovam Mendonça	Christovam Mendonça Filho
Antonio Carlos Bloise	Antonio Carlos Bloise Júnior
José Pedro Donoso	Jose Pedro Donoso Gonzalez
José Schneider	José Fabián Schneider
A. Kinoshita	Angela Mitie Otta Kinoshita

Confrontando-se dados de periódicos da SciELO com os currículos dos pesquisadores autores dos respectivos artigos, identificou-se na categoria dos campos com Autonomia Total, no preenchimento: erros de digitação, o uso do idioma inglês (quando o sistema, maiormente adota a língua portuguesa), e até mesmo a ausência ou troca nos títulos. Essas falhas comprometem o sistema: algumas sugestões para correções destas falhas serão apontadas ao longo do texto, pois dizem respeito também a outras categorias de campos.

4.2.2 ANÁLISE DOS CAMPOS COM AUTONOMIA PARCIAL

Conforme explicado, trata-se de campos inicialmente sem opções (similar aos campos com Autonomia Total), e cada novo termo cadastrado é armazenado no sistema. Para a análise dos campos com Autonomia Parcial foram considerados aspectos de sinonímia e homonímia, que são representações lingüísticas diferentes para objetos iguais ou similares, o que demonstra a natureza semântica destes campos.

Para a análise foram confrontadas as palavras-chave cadastradas pelos autores na **PL** com as palavras-chave registradas nos artigos publicados nos periódicos disponíveis na SciELO[50]. A análise segue o modelo a seguir:

[50] A SciELO exige que os periódicos usem termos de acordo com normas que sejam compatíveis com padrões internacionais de bases de dados. Cabe ao periódico estabelecer suas normas editoriais, desde que as mesmas se enquadrem nos critérios da SciELO.

MODELO
Exemplo: Nome do Periódico

PERIÓDICO	AUTOR
Palavras-chave que constam no artigo.	Palavras-chave cadastradas pelo autor na **PL**

Inconsistências: **Serão indicados os elementos que desfavorecem a consistência de um sistema de informação[51].**

Exemplo 1: Ciência da Informação

PERIÓDICO	AUTOR
Organização do conhecimento	Organização do conhecimento
Ciberespaço	
Mecanismos de busca	Mecanismos de busca
	Rizoma
	Tecnologias da Informação

Inconsistências: Dispersão de termos.

Exemplo 2: Delta

PERIÓDICO	AUTOR
escrita	Escrita
escola	
chat	
internet	Internet
	Comunicacao eletronica
	Letramento Escolar
	letramento digital

Inconsistência: Dispersão de termos.

Exemplo 3: Arquivo Brasileiro de Medicina Veterinária e Zootecnia

PERIÓDICO	AUTOR 1	AUTOR 2	AUTOR 3	AUTOR 4
cão				
fluorquinolona	fluorquinolona			
intoxicação	intoxicação			intoxicação
choque	choque			
enrofloxacina	ENROFLOXACINA			enrofloxacina
	cães			cães
		Clinica de pequenos animais		
		Clínica		
			enrofloxacin	
			canine	
			fluoroquinolone	
			overdose	
			shock	

Inconsistências: Uso do plural, dispersão de termos, idioma diferente do português.

[51] Ainda que no periódico não tenha sido adotada a forma no singular, foi considerada inconsistência o uso do plural em razão de a normalização gramatical preconizar a forma no singular.

Exemplo 4: Memórias do Instituto Oswaldo Cruz[52]

PERIÓDICO	AUTOR 1	AUTOR 2	AUTOR 3	AUTOR 4	AUTOR 5
Taenia solium cysticercus antibodies		Taenia solium cysticercus antibodies			
enzyme linked immunoabsorbent assay					
Immunoblot - IgE - Brazil		immunoblot IgE Brazil			
	immunoblot				imunoblot
	Taenia solium				Taenia solium
	ELISA				ELISA
		immunoabsorbent assay			
		enzyme linked			
			Cysticercosis	cysticercosis	
				Pontal do Paranapanema	
				Teodoro Sampaio	
				Taeniosis	
					Taenia saginata

Inconsistências: Dispersão de termos, idioma diferente do português[53].

Exemplo 5: Brazilian Journal of Physics

PERIÓDICO	AUTOR 1	AUTOR 2	AUTOR 3	AUTOR 4
Magnetic resonance imaging	Magnetic Ressonance Imaging			
MRI	MRI			MRI
		Tempos de Relaxação		
		Imagem por Ressonância Magnética		
		Fígado		
		Seio		
			Relaxometry	
			Magnetic Ressonance	
			Echo Time	

Inconsistências: Dispersão de termos, idioma diferente do português.

Exemplo 6: Brazilian Journal of Chemical Engineering

PERIÓDICO	AUTOR 1	AUTOR 2	AUTOR 3
Electrostatic charges			electrostatic charges
Charges measurement			
Aerosol particles			aerosol particles
	Cargas Eletrostaticas		
	Eletromobilidade		
	Aerossois		
		separation	
		particles	
		electrostatic	
		electric field	

Inconsistências: Dispersão de termos, uso do plural, idioma diferente do português.

[52] Nos exemplos 4, 5 e 6, os periódicos não publicam palavras-chave em português. No entanto, no cadastramento de palavras-chave pelo autor, na PL, há variação quanto à língua adotada.
[53] Na **PL** não há recurso que estabeleça a compatibilidade de termos para idiomas diferentes.

Exemplo 7: Arquivos Brasileiros de Cardiologia

PERIÓDICO	AUTOR 1	AUTOR 2
exercício	Exercício	
insuficiência cardíaca congestiva	insuficiência cardíaca congestiva	insuficiência cardíaca congestiva
idade	Idade	
		Ventilação Pulmonar
		teste de esforço
		Consumo de Oxigênio
		Fatores etários
		Limiar Anaeróbio

Inconsistência: Dispersão de termos.

Os exemplos tornam evidente que pode haver o preenchimento inadequado de palavras-chave devido à abertura na **PL** para o preenchimento das palavras-chaves. Nos exemplos 1 e 2 um fato chama a atenção: ambos foram publicados por um único autor. Esse autor é responsável por indicar as palavras-chave em seu artigo. Porém, ao cadastrar a publicação desse mesmo artigo em seu currículo, o autor deixou de usar palavras-chave que ele indicou para o artigo e ainda acrescentou outros que não foram indicados para o artigo.

É patente a adoção de termos no plural. Sabe-se que, para fins documentários, a normalização gramatical (que será discutida adiante) é preconizada para sistemas de informação, prevendo evitar divergências na grafia das palavras. Diferenças na grafia ocasionadas pelo uso do plural/singular não interferem em alguns sistemas de buscas que identificam a ausência da letra “S” no final da palavra. Mas, no caso do exemplo 3, as palavras “CÃO” e “CÃES” são compreendidas como representações de objetos diferentes, quando na verdade trata-se apenas de numeral.

Uma situação identificada na categoria dos campos com Autonomia Parcial, que ocorre também nos campos com Autonomia Total, é o uso de termos em língua diferente do português. Um dos motivos que conduz os autores a cadastrarem as palavras em outro idioma é que os artigos, mesmo publicados no Brasil, estão em outra língua. É previsível que os autores usem a língua adotada na publicação, porém, há um fato curioso: os autores utilizam a mesma língua, mas não necessariamente repetem as mesmas palavras-chave usadas no artigo. Como pode ser verificado nos exemplos 4,5 e 6, há palavras-chave diferentes das indicadas nos artigos que foram cadastradas também em inglês.

A inserção de palavras na **PL** em idiomas diferentes do português prejudica o processo de busca e recuperação da informação. Deveria haver um dicionário que compatibilizasse as palavras para que o sistema fosse capaz de interpretar a paridade entre termos em português e inglês, identificando as correspondências entre as duas línguas. Ou então, seguir uma opção: o usuário seria orientado - no processo de preenchimento - a utilizar somente palavras-chave em português, ou, deveria haver campos que dessem a opção para o preenchimento em mais de uma língua.

Utilizar mais de um idioma para criar palavras-chave não seria algo novo para os pesquisadores, pois as normas para publicações científicas já exigem resumos e palavras-chave em pelo menos uma língua diferente (normalmente em idioma inglês). Tal procedimento poderia ser adotado também para o campo de título na **PL**.

A importância do idioma adotado no preenchimento de campos acentua-se ainda mais em razão de haver acordos internacionais envolvendo a **PL**. Iniciativas nesse sentido já foram feitas. Segundo Santana et al (2001, p.48), em 2000, a BIREME, a OPAS (Organização Pan-Americana da Saúde), a OMS (Organização Mundial da Saúde) e o CNPq acordaram o projeto cooperativo para estabelecer enlaces entre a SciELO e a base de dados de currículos da **PL** mantida pelo CNPq.

De acordo com Rios e Santana (2001) foi apresentado a organizações nacionais de C&T do Chile, da Colômbia, da Venezuela, do México e de Cuba o Sistema de Currículos Vitae em Ciências da Saúde. Além da aprovação deste sistema, foi solicitado que não se considerassem apenas as áreas da saúde, redundando na elaboração do Sistema de Currículos Vitae Latino-Americano e do Caribe (Sistema CvLAC)[54].

Entre as inconsistências em um sistema de informação, a dispersão de termos é uma das mais comprometedoras. Os problemas acarretados envolvem aspectos de natureza tecnológica (como a agilidade no sistema devido à extensa lista de termos no banco de dados, o que influencia na rapidez da resposta do sistema) e também elementos relacionados a princípios

[54] Segundo Rios e Santana (2001) o CvLAC é um espaço comum de integração e intercambio de informação de currículos dos atores da C&T de países da América Latina e Caribe (Brasil, Colômbia, Cuba, Chile, México e Venezuela).

de organização e tratamento da informação, redundando em problemas na Recuperação da Informação.

Não será aprofundada a discussão das deficiências de ordem tecnológica, porém ressalta-se que há uma relação direta entre a quantidade de termos armazenada no banco de dados da **PL** e a rapidez nas respostas de busca. O princípio é simples: quanto mais registros diferentes de palavras-chave houver no sistema, maior o tempo para processar buscas num índice de termos.

A velocidade de resposta numa busca é um aspecto importante, mas há que se considerar nesse o avanço das TICs, pois a velocidade de processamento dos sistemas é cada vez mais influenciada pelas infra-estruturas tecnológicas, razão pela qual as questões de processamento lógico do sistema (por exemplo, conforme acima apontado, diferenças entre termos no singular e plural) tendem a não ser mais tão prejudiciais como já foram um dia.

Mais condizentes com as propostas desta pesquisa são as discussões relativas à organização e tratamento da informação, e o alto índice de dispersão deixa claro que há uma necessidade de se solucionar as inconsistências verificadas.

Para este estudo entender-se-á que a "dispersão" resulta da diversidade de palavras-chave usadas para representar uma dada produção cientifica, ou seja, a ausência de controle na inserção de palavras-chave na **PL,** por parte dos autores de cada artigo. Essa diversidade acarreta a "pulverização" da informação[55].

Tradicionalmente, as palavras-chave usadas em documentos servem como representações temáticas dos próprios documentos. Tais representações não almejam completar a mensagem, e sim oferecer um recurso auxiliar para recuperação da informação. Na **PL**, as palavras-chave da produção científica tanto podem ser utilizadas no processo de recuperação dos currículos, como também proporcionar estudos métricos da produção científica brasileira e respectivos indicadores de C&T.

---

[55] Na PL existe o limite de seis palavras-chave para cada artigo, este número não destoa da média adotada pelos periódicos, isso indica que o aspecto quantitativo das palavras-chave é um fator secundário, mais urgente é uma atenção à dispersão e falta de orientação para o preenchimento.

Tanto para fins de recuperação da informação como para estudos métricos, é importante que - além do planejamento do sistema para essas finalidades – a base de dados de currículos seja alimentada a partir de uma orientação voltada a esses propósitos. Mas os exemplos demonstraram o contrário. A característica aberta da **PL** permite que o preenchimento dos campos seja realizado à mercê da percepção que os usuários alimentadores têm do seu funcionamento ou dos objetivos que eles perseguem no momento do preenchimento e dos objetivos da própria **PL**.

No exemplo 7 percebe-se que um dos autores utilizou as mesmas palavras-chave do periódico, diferentemente do outro autor que adotou outros termos segundo a percepção dele do que seria melhor. No mesmo exemplo observa-se que, enquanto o periódico (por indicação dos autores) adotou o termo IDADE, o autor 2 optou por FATORES ETÁRIOS. Não se trata exatamente de um erro, porém de um procedimento que diverge das orientações comuns à organização da informação segundo uma visão documentária, nesse caso a dispersão de um conceito entre dois termos quase-sinônimos.

Era de se esperar inconsistências relativas à sinonímia e/ou homonímia nos campos de palavras-chave em sistemas abertos. No caso da **PL**, para o preenchimento de palavras-chave, não há esclarecimentos a respeito do uso dos termos; assim, é improvável que - desconhecendo princípios de organização da informação – os usuários se preocupassem com questões de natureza documentária.

A atual quantidade de registros na **PL** torna possível aos seus administradores fazer um levantamento (a partir da base de dados do sistema) representativo dos termos mais utilizados no campo das palavras-chave. Se esse levantamento fosse realizado, é provável que houvesse viabilidade técnica para estratificar os termos segundo áreas de conhecimento dos currículos. A partir de uma relação dos termos mais adotados em cada área é viável implementar um recurso que auxiliasse o usuário a preencher os campos, sugerindo os termos mais adotados por seus pares.

Ressalta-se que a finalidade não seria eliminar a especificidade, ou seja, aquilo que por ser menos freqüente pudesse ser mais informacional: o intuito é apresentar uma opção de grafia a partir das primeiras letras do termo que o

usuário estivesse cadastrando. A tecnologia adotada pelo *Google Suggest*[56] segue esse princípio (Figura 15):

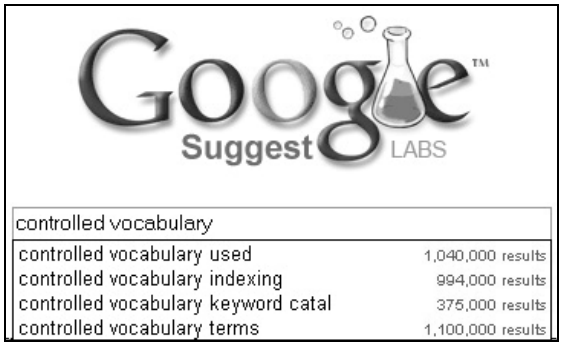

Figura 15 – Google Suggest

No entanto, essa não é uma solução para as atuais inconsistências da **PL** no que diz respeito à organização da informação. Serviria somente como uma forma de orientação do sistema para o preenchimento dos campos, considerando-se que atualmente não há, com exceção da lista de termos criada pelo próprio usuário (que também pode conter sinonímias, formas gramaticais diferentes, etc.), indicação de quais palavras-chave o usuário poderia adotar. Isso permitiria ao usuário ter uma noção (quantitativa) dos termos mais adotados por seus pares. Na forma atual, o usuário, no momento do preenchimento, visualiza apenas os termos que ele próprio cadastrou.

É necessário ressaltar que nas produções interdisciplinares, ou seja, elaboradas colaborativamente por autores de distintas áreas do conhecimento, as sugestões do sistema seriam baseadas em um conjunto de termos usuais as respectivas áreas de conhecimento dos autores, ou mesmo a partir de áreas que o próprio usuário pudesse definir previamente. Desta forma, quando o usuário inserir as primeiras letras do termo pretendido para o preenchimento, o sistema apresentará opções baseadas na similaridade sintática das palavras, ou seja, a partir da coincidência da grafia dos termos.

No entanto, não é possível ignorar que a identificação da(s) área(s) de conhecimento na(s) qual(ais) o(s) autor(es) atua(m) é muito difícil, particularmente no caso de áreas interdisciplinares ou recém-configuradas.

[56] Serviço em fase de teste do laboratório da Empresa Google. Enquanto se digita a palavra no campo de busca são oferecidas sugestões em tempo real pelo sistema. Assim, à medida que novas letras são inseridas, as opções podem se modificar. Ao lado de cada sugestão consta a quantidade de resultados referentes a respectiva sugestão. Endereço: http://www.google.com/webhp?complete=1&hl=en.

É visível que as inconsistências na categoria dos campos com Autonomia Parcial da **PL** são prejudiciais à RI. Mas, a inconsistência que gera dispersão é desfavorável principalmente às análises conjunturais dos currículos, comprometendo a desejada formulação de indicadores de C&T a partir de dados da **PL**.

Para que um repositório de informações em C&T seja capaz de gerar indicadores confiáveis é necessária a padronização tanto dos dados bibliográficos quanto temáticos. Na opinião de Kobashi e Santos (2007, p.5), os acervos da produção científica brasileira são dispersos, pouco padronizados e apresentam inconsistências em quase todos os campos dos registros bibliográficos. Essa opinião revela a dificuldade dos autores para desenvolver um estudo bibliométrico a partir de teses e dissertações nas áreas da Ciência da Informação e Energia Nuclear. Foi necessário reformatar grande parte dos dados para se chegar a um maior grau de homogeneização, ou seja, alcançar um nível de generalidade capaz de representar classes com temáticas afins.

Notou-se, nos exemplos analisados da **PL,** que a representação usual dos pesquisadores segue uma tendência para uso de uma linguagem natural. As relações entre palavras-chave de autores e periódico e vice-versa, demonstraram que procedimentos requeridos em sistemas fechados de informação são poucos usuais. Para uma análise aprofundada dos termos exemplificados seria necessário o domínio das áreas de conhecimento dos artigos publicados. Contudo, uma compreensão superficial sobre algumas temáticas já é o suficiente para perceber que Ressonância Magnética por Imagem é um tipo de ressonância magnética (exemplo 5), canino e cães são sinônimos (exemplo 3), chat é hipônimo de comunicação eletrônica (exemplo 2).

É importante entender que a linguagem controlada busca, justamente, reduzir as variações semânticas e sintáticas de uma linguagem natural. Neste caso, são importantes as sinonímias, homonímias, além das opções de grafias, e ainda a designação de um termo único como portador de um único conceito, definindo-se assim palavras preferidas e não-preferidas para representar a informação. A funcionalidade e êxito da linguagem controlada – ou linguagem documentária - limita-se a ambientes de informação, com o objetivo de organizar e recuperar a informação; são linguagens construídas e, por isso,

consideradas artificiais, não tendo aplicabilidade em outros ambientes e situações.

No que diz respeito ao controle de termos, a categoria dos campos com Autonomia Parcial da **PL** é um pouco menos crítica que a dos campos com Autonomia Total**,** pois possibilita o re-uso de termos já existentes no currículo do pesquisador. Novas e desnecessárias palavras-chave podem ser evitadas graças ao fato de o sistema gerar uma lista de termos, que poderá ser consultada quando da atualização de um currículo, o que pode minimizar a inclusão de sinônimos, hipônimos, e plurais. No entanto, esta possibilidade não basta para que o sistema proporcione uma recuperação da informação eficaz e tampouco seja capaz de produzir, a partir de sua base de dados, indicadores em C&T consistentes.

Interessantes reflexões de Kobashi e Santos (2007) dizem respeito a essas problemáticas da **PL** concernentes à produção de indicadores. Os autores explicam que os dados temáticos necessários para a produção de indicadores não podem ser tratados segundo as mesmas políticas de indexação para fins de RI. A especificidade é o princípio básico aplicado na indexação para recuperação, que objetiva discriminar informação por meio da criação de classes constituídas por uma quantidade manejável de registros bibliográficos. Tal procedimento desfavorece estudos bibliométricos.

Em geral, os termos utilizados na indexação para recuperação proporcionam grande quantidade de classes de baixa freqüência, resultando em núcleo reduzido e alta dispersão. Por outro lado, é preciso cautela na reformatação para que a substituição de termos específicos por níveis mais genéricos não gere classes com freqüências muito altas, pois freqüências altas tendem a não apresentar significados.

Há características da **PL** que indicam problemas de planejamento e operação relacionados às suas finalidades. No que tange ao planejamento, observou-se que a concepção, apesar de (supostamente) orientada ao desenvolvimento de um sistema para recuperação e geração de indicadores para Gestão de C&T, não previu - em campos importantes como os das

palavras-chave - meios necessários a estes fins como, por exemplo, um vocabulário controlado ou uma arvore hierárquica de termos[57].

Com relação ao funcionamento da **PL**, há dois aspectos: o preenchimento dos currículos pelo usuário e o uso efetivo do sistema para buscas ou utilização dos dados da base. É no funcionamento que a ausência de controle inicia o processo que acarretará deficiências no sistema, pois é no preenchimento livre dos campos que os usuários inserem dados no sistema que apresentarão inconsistências, como as mostradas nos exemplos desta pesquisa.

Segundo seus desenvolvedores (GRUPO STELLA, 2007), a **PL** seguiu um modelo em que os usuários "são produtores e multiplicadores de conhecimento (pesquisadores, docentes, estudantes, grupos de pesquisa, etc.)". Assim, o conjunto que configura a **PL** é baseado num princípio denominado de "regras de negócio dos sistemas" onde cada um dos usuários utiliza e gera a informação que conformará o sistema. No entanto, ao optar por seguir a chamada regra de negócio dos sistemas, o desenvolvimento da **PL** priorizou a economia de custos, abrindo mão da sua qualidade.

O debate sobre "regras de negócio dos sistemas" está relacionado com os novos modelos de serviços da Web, abertos à participação dos usuários para o compartilhamento de serviços e informações. Na percepção de Catarino e Baptista (2007), trata-se de um novo paradigma para a organização dos conteúdos de recursos digitais na Web designados, genericamente, de *folksonomias*. Já na visão de Noruzi (2007), a folksonomia corresponde a uma taxonomia auto-gerada (no original user-generated) pelo usuário para que ele – o usuário - possa categorizar e recuperar conteúdos da Web a partir de etiquetas denominadas "tags". De acordo com esse autor, as tags podem contribuir para a melhoria dos sistemas de busca da internet, em razão de os conteúdos categorizados formarem um vocabulário compartilhável entre usuários.

É precipitado considerar as folksonomias como um novo paradigma, pois a criação de etiquetas (tags) de marcação para conteúdos na Internet, por enquanto não configura, sob o ponto de vista da organização da informação,

[57] Mais discussões sobre a hierarquização na Subseção 3.3.

um paradigma que possa ser expandido para todo e qualquer contexto informacional.

Princípio semelhante já existe há mais de dez anos nas páginas geradas em HTML através de comandos que permitem ao criador do documento definir palavras-chave e resumos para cada página criada, além de descrever textualmente imagens. Abaixo (Figura 16) é mostrado um exemplo destes recursos:

Função	Comando em HTML
Incluir palavras-chave	meta name= "keywords" content=" Folksonomia; Etiquetagem social; Etiquetagem colaborativa"
Incluir resumo	meta name="description" content= "Apresenta novo conceito para a organização dos recursos digitais na Web: a folksonomia."
Descrever imagens	img border="0" src="papa.jpg" width="290" height="286" alt= "Papa Bento XVI no Brasil "

Figura 16 – Exemplos de recursos em HTML

Esses comandos em HTML não interferem no conteúdo apresentado na página, eles são visíveis somente aos sistemas de buscas da Web como o Google, que os utiliza para “indexar” as páginas em seus índices. Evidentemente que a diferença principal entre o uso desses comandos e os serviços da folksonomia está no papel que os usuários assumem. Nos comandos HTML os usuários (excetuando-se os criadores da página) têm papel passivo na representação dos conteúdos, enquanto que nas folksonomias cabe a eles a função de representar os conteúdos segundo seus interesses.

Convergindo a discussão para o foco desta pesquisa, acredita-se que as folksonomias, enquanto recurso para organização da ICT, apresentam-se ainda em fase de desenvolvimento bastante incipiente, sendo necessário maior amadurecimento e consolidação de seus conceitos. As folksonomias são relacionadas à indexação pelo usuário, ao passo que os vocabulários controlados estão associados à indexação voltada para sistemas de informação.

No caso da **PL,** o processo de representação através de termos se assemelha mais aos recursos do HTML do que das etiquetagens das folksonomias, pois na **PL** os pesquisadores que preenchem os currículos é que definem as palavras que poderão servir, por exemplo, para fins de recuperação

ou estudos (ao invés da indexação pelo usuário). Além disso, as fragilidades das folksonomias apontadas por Catarino e Baptista (2007) indicam que os problemas já identificados nesses sistemas são tão comprometedores como os da **PL.** Cita-se alguns:

- a liberdade de atribuição de etiquetas faz com que haja pouca precisão na recuperação da informação;
- as palavras-chave atribuídas pelos usuários são frequentemente ambíguas e inexatas;
- por enquanto, há pouco ou nenhum controle de sinônimos ou homônimos, e não há regras de indexação;
- quanto aos aspectos semânticos da classificação, os sistemas de etiquetagem precisam resolver problemas que são inerentes ao processo de criação de relações semânticas entre palavras, como a polissemia e a sinonímia.

Noruzi (2006) ressalta problemas similares nas folksonomias, ressaltando mais especificamente quatro: polissemia, sinonímia, plural e especificidade. Entretanto, apesar de suas desvantagens, Noruzi acredita que as folksonomias representam uma mudança na metodologia de classificação a partir da distribuição e descentralização de tarefas. O referido autor destaca que as folksonomias removem todo o conceito de hierarquia dos esquemas classificatórios para a organização do conhecimento. Contudo, Noruzi está ciente da necessidade de evolução das folksonomias e também da urgência do controle de vocabulário neste novo recurso voltado para a organização da informação.

4.2.3 ANÁLISES DOS CAMPOS SEM AUTONOMIA

Nos campos sem autonomia o sistema oferece um conjunto de opções pré-cadastradas ao usuário. Em campos como “Áreas do Conhecimento”, “Setores de Atividade” e “Título do Periódico”, por exemplo, deve-se preferencialmente cadastrar itens pertinentes por consulta aos itens pré-cadastrados no sistema. Porém, é facultado incluir novos itens que não constem desse conjunto de opções. Na página de busca avançada da **PL**[58] há

[58] http://buscatextual.cnpq.br/buscatextual/index.jsp acessada em 20/08/2007.

filtros[59] que facilitam e refinam o processo de busca. Ao aplicar um ou mais filtros, o usuário aumenta as chances de harmonizar sua estratégia de busca com os registros da base, pois as opções oferecidas pelos filtros reproduzem as mesmas oferecidas aos usuários no preenchimento dos campos Sem Autonomia. Exemplo: No preenchimento, para cadastrar sua "Área de Atuação", o usuário deve escolher uma opção a partir de uma lista pré-definida. Já no processo de busca, é oferecido ao usuário em um dos filtros a mesma lista de Áreas de Atuação disponível para preenchimento de currículos.

Para análise dos campos Sem Autonomia foi observado o campo "Áreas do Conhecimento". Contudo, diferentemente das categorias anteriores, não foram feitas comparações a artigos de periódicos da SciELO. Foram analisadas somente as representações de Áreas de Conhecimento que os autores fizeram dos artigos disponíveis na SciELO. O formato dos exemplos segue o modelo:

MODELO
Periódico: Título do periódico
Área do Conhecimento: Área do conhecimento na qual o periódico se enquadra na SciELO.
Título: Título do artigo a qual o(s) autor(es) faz(em) referência na **PL**.

Autor 1
Área do conhecimento indicada[60] pelo autor, seguindo a seguinte hierarquia: Grande Área / Área / Subárea / Especialidade

Observação: Será usado o termo OBSERVAÇÃO para indicar as particularidades de cada exemplo que podem proporcionar problemas quanto à recuperação e uso da informação.

EXEMPLO 1
Periódico: Ciência da Informação
Área do Conhecimento: Ciências Sociais Aplicadas
Título: Redes neurais e sua aplicação em sistemas de recuperação de informação

Autor 1
Ciências Sociais Aplicadas / Ciência da Informação.

Observação: Representação genérica da produção (indicação apenas da Área de Conhecimento).

EXEMPLO 2
Periódico: DELTA
Área do Conhecimento: Letras e Artes
Título: Análise de conteúdo e análise do discurso: o lingüístico e seu entorno

Autor 2	**Autor 2**
Lingüística, Letras e Artes/Lingüística.	
Lingüística, Letras e Artes/Lingüística/Lingüística Aplicada.	
Lingüística, Letras e Artes/Lingüística/Lingüística Aplicada/Especialidade: Análise do Discurso	
	Lingüística, Letras e Artes
	Ciências Humanas/Educação

Observação: Representação genérica da produção (indicação apenas da Área de Conhecimento); Relações partitivas diferentes; Indicação diferente de Grandes Áreas.

[59] Pesquisadores com algum tipo de bolsa, Formação Acadêmica, Área de Atuação, Atividades de Orientação, Áreas ou Setores da Produção em C&T, Atividade Profissional, e Presença no Diretório de Grupos de pesquisa.
[60] O autor pode indicar mais de uma área.

EXEMPLO 3

Periódico: Arquivo Brasileiro de Medicina Veterinária e Zootecnia
Área do Conhecimento: Ciências Agrárias
Título: Aplicação da técnica de PCR na detecção de Yersinia enterocolitica em suínos abatidos sem inspeção

Autor 1	Autor 2	Autor 3
Ciências Agrárias/Medicina Veterinária/Inspeção de Produtos de Origem Animal.	Ciências Agrárias/Medicina Veterinária/Inspeção de Produtos de Origem Animal.	
Ciências da Saúde/Saúde Coletiva / Saúde Pública.		
Ciências Agrárias/Medicina Veterinária/Medicina Veterinária Preventiva.		
	Ciências Agrárias/Medicina Veterinária /Saúde Pública Veterinária.	
	Ciências Agrárias/Ciência e Tecnologia de Alimentos /Ciência de Alimentos / Especialidade: Avaliação e Controle de Qualidade de Alimentos.	
		Ciências Biológicas / Microbiologia.
		Ciências Biológicas /Microbiologia / Microbiologia Aplicada

Observação: Relações partitivas diferentes; Indicação diferente de Grandes Áreas, Representação genérica da produção (indicação apenas da Área de Conhecimento).

EXEMPLO 4

Periódico: Memórias do Instituto Oswaldo Cruz
Área do Conhecimento: Ciências Biológicas
Título: Taeniosis-cysticercosis complex in individuals of a peasants' settlement (Teodoro Sampaio, Pontal of Paranapanema, SP, Brazil)

Autor 1	Autor 2	Autor 3	Autor 4	Autor 5
Ciências Biológicas/Imunologia.				
Ciências Biológicas/ Imunologia/Imunologia Aplicada.				
Ciências Biológicas/Parasitologia.				
	Ciências da Saúde/Farmácia			
		Ciências da Saúde/Farmácia/Análises Clínicas.		
		Ciências Biológicas/Parasitologia/Helmintologia de Parasitos.		
			Ciências da Saúde/Farmácia /Análises Clinicas Imunologia.	
				Ciências da Saúde/Medicina

Observação: Relações partitivas diferentes; Indicação diferente de Grandes Áreas; Representação genérica da produção (indicação apenas da Área de Conhecimento).

EXEMPLO 5

Periódico: Brazilian Journal of Chemical Engineering
Área do Conhecimento: Engenharias
Título: Oxidation of limonene catalyzed by Metal(Salen) complexe

Autor 1	Autor 2	Autor 3
Ciências Exatas e da Terra /Química /Físico-Química /Cinética Química e Catálise.		
	Engenharias /Engenharia Química.	
		Engenharias /Engenharia Química /Processos Químicos /Termodinâmica
		Engenharias /Engenharia Química /Processos Industriais de Engenharia Química /Processos Orgânicos.
		Engenharias /Engenharia Química /Operações Industriais e Equipamentos para Engenharia Química /Reatores Químicos.
		Engenharias /Engenharia Química /Tecnologia Química /Produtos Naturais.

Observação: Relações partitivas diferentes; Indicação diferente de Grandes Áreas; Representação genérica da produção (indicação apenas da Área de Conhecimento).

EXEMPLO 6

Periódico: Brazilian Journal of Physics
Área do Conhecimento: Ciências Exatas e da Terra
Título: Electron spin resonance dating of shells from the sambaqui (shell mound) Capelinha, São Paulo, Brazil

Autor 1	Autor 2	Autor 3
Ciências Exatas e da Terra/Física.		
Ciências Exatas e da Terra/Física/Física Médica e Biológica.		
	Ciências Exatas e da Terra/Física/Física Aplicada a Medicina e Biologia.	
		Ciências Humanas/Arqueologia
		Ciências Exatas e da Terra/Física/Física das Partículas Elementares e Campos.
		Ciências Humanas/Arqueologia/ Arqueologia Pré-Histórica

Observação: Representação genérica da produção (indicação apenas da Área de Conhecimento); Relações partitivas diferentes; Indicação diferente de Grandes Áreas.

EXEMPLO 7
Periódico: Arquivos Brasileiros de Cardiologia
Área do Conhecimento: Ciências da Saúde
Título: Respostas cardiopulmonares ao exercício em pacientes com insuficiência cardíaca congestiva de diferentes faixas etária

Autor 1	Autor 2	Autor 3
Ciências da Saúde/Medicina/SubClínica Médica/ Cardiologia.		
	Ciências da Saúde / Medicina.	Ciências da Saúde / Medicina.
	Ciências da Saúde / Educação Física.	
	Ciências da Saúde / Fisioterapia e Terapia Ocupacional	
		Ciências Biológicas / Fisiologia.

Observação: Relações partitivas diferentes; Indicação diferente de Grandes Áreas; Representação genérica da produção (indicação apenas da Área de Conhecimento).

EXEMPLO 8
Periódico: Estudos Avançados
Área do Conhecimento: Ciências Humanas
Título: A universidade primeira do Brasil: entre intelligentsia, padrão internacional e inclusão social.

Autor 1
Ciências Humanas / Educação / Educação Superior.
Ciências Humanas / Ciência Política / Política Científica e Tecnológica.
Ciências Humanas / Ciência Política / Política Científica e Tecnológica / Política de Ciência e Tecnologia.

Observação: Relações partitivas diferentes; Indicação diferente de Áreas do Conhecimento.

Dos três tipos de campos para preenchimento da **PL**, os campos Sem Autonomia constituem, sem dúvida, a opção mais restritiva no que diz respeito à liberdade de inserção do usuário. Ainda que exista a possibilidade de se incluir palavras diferentes das listadas, em nenhum currículo foi identificada alguma inclusão.

Nos campos da **PL** "Áreas do Conhecimento" e "Setores de Atividades" as opções seguem uma estrutura hierárquica. As estruturas hierárquicas permitem uma visualização de níveis mais genéricos para os mais específicos. As Áreas de Conhecimento estão baseadas na Tabela de Áreas de Conhecimento do CNPq, enquanto que o campo Setores de Atividades dispõe uma classificação própria do sistema, que indica setores econômicos e sociais relacionados ao trabalho desenvolvido.

É visível nos exemplos que as diferenças entre as indicações dos autores variaram não somente em razão da relação geral/específico. No exemplo 1 há somente uma indicação genérica limitada a uma grande área do conhecimento. No exemplo 2, enquanto um autor optou por representar explorando o aspecto da especificidade, o outro preferiu indicar duas grandes

áreas do conhecimento (uma delas inclusive não foi mencionada pelo anterior).

No exemplo 3 há indicações diferentes para as grandes áreas do conhecimento (Ciências Agrárias, Ciências da Saúde e Ciências Biológicas) havendo apenas uma indicação coincidente entre os autores analisados. No exemplo 4, apesar de existirem concordâncias entre autores em alguns níveis das relações partitivas, chama a atenção o fato de não haver nenhuma representação exatamente igual entre eles.

No exemplo 5 ocorre um fato similar ao exemplo 4: há concordâncias entre autores em alguns níveis das relações partitivas, mas não há nenhuma representação exatamente igual entre eles. Tal fato se repete no exemplo 6, porém chama a atenção o registro de duas áreas bastante distintas: a Física e a Arqueologia. O exemplo 7 traz maiores distinções no que diz respeito às áreas do conhecimento. Nos três autores analisados há o registro da produção como na área da Medicina, da Educação Física, da Fisioterapia e Terapia Educacional e da Fisiologia. O exemplo 8 indica somente uma grande área do conhecimento. As relações partitivas subseqüentes, por serem coerentes, não permitem a detecção prévia de problemas de consistência que possam comprometer a recuperação.

Além da diferença de ordem hierárquica, identificaram-se discordâncias entre autores de um mesmo artigo quanto à escolha das áreas. No exemplo 2, um dos autores indicou a área Ciências Humanas/Educação, quando o artigo parece ser mais focado na área da Lingüística. Isso não se configura, necessariamente, como uma falha, mas sim, a percepção pessoal dos pesquisadores devida, talvez, a suas respectivas áreas de atuação. De qualquer forma, é patente que o consenso absoluto não é comum entre os autores quanto à escolha das Áreas de Conhecimento no qual se insere o artigo e esse fato pode acarretar inconsistências no uso dos dados como indicadores científicos.

O quadro a seguir (Figura 17) demonstra essas diferenças quando o artigo apresenta co-autoria[61]:

Exemplo	Grande Área	Área do Conhecimento
2	Lingüística, Letras e Artes/ Ciências Humanas	Lingüística Educação
3	Ciências Agrárias/ Ciências da Saúde/ Ciências Agrárias/ Ciências Biológicas/	Medicina Veterinária Saúde Coletiva Ciência e Tecnologia de Alimentos Microbiologia
4	Ciências Biológicas/ Ciências Biológicas/ Ciências da Saúde/ Ciências da Saúde/	Imunologia Parasitologia Farmácia Medicina
5	Ciências Exatas e da Terra/ Engenharias/	Química Engenharia Química
6	Ciências Exatas e da Terra/ Ciências Humanas/	Física Arqueologia
7	Ciências da Saúde/ Ciências da Saúde/ Ciências da Saúde/ Ciências Biológicas/	Medicina Educação Física Fisioterapia e Terapia Ocupacional Fisiologia

Figura 17 - Identificação de Áreas de Conhecimento em artigos com co-autoria

São perceptíveis as diferenças nas indicações das Grandes Áreas e, principalmente, Áreas de Conhecimento. Do ponto de vista do uso das informações da **PL** para a Gestão em C&T isso pode significar que: há um conjunto fragmentado de dados pouco informativo para indicar comportamentos no âmbito da produção científica brasileira, ou então, quando pesquisadores de diferentes áreas produzem conjuntamente acentua-se o caráter multidisciplinar/interdisciplinar de co-autorias. Em ambas as situações, a interpretação adequada dos dados exigirá uma rigorosa compreensão dessas nuances.

Quanto à **RI**, o uso dos campos Sem Autonomia na **PL** é relativamente bem explorado para fins de busca de currículos. É oferecida a possibilidade de busca pela produção de acordo com as Áreas de Conhecimento. Como a estratégia de busca é formulada a partir de uma lista controlada, torna-se mais fácil estabelecer coincidências entre os termos definidos pelos usuários com os existentes na base do sistema.

Por fim, os motivos que conduzem os usuários a preencherem os campos Sem Autonomia com termos genéricos ou então com termos que representam coisas distintas (ex: FÍSICA – ARQUEOLOGIA; LINGÜÍSTICA – EDUCAÇÃO; MEDICINA – EDUCAÇÃO FÍSICA) requerem investigações mais

[61] Os Exemplos 1 e 8 foram desconsiderados por terem apenas um autor.

apropriadas a esse fim e, para tanto, são necessários estudos de usuários focados na representação da informação. Não se deve desconsiderar que a raiz do problema pode também estar presente no recurso que é oferecido ao usuário. No caso específico do exemplo explorado, a árvore hierárquica talvez seja insuficiente para representar de forma exaustiva a diversidade de Áreas de Conhecimento.

O campo Áreas de Conhecimento é preenchido para cada atividade ou produção do pesquisador. Desta forma, um agrônomo que atua no segmento de defesa fitossatinária, provavelmente terá um currículo com atividades e produções direcionadas ao setor de fitossanidade; entretanto, é totalmente possível que, se for necessário, ele cadastre em seu currículo uma palestra na área da saúde pública.

Um fato curioso chama a atenção: a indicação da Área de Conhecimento por parte dos autores dos artigos do periódico Ciência da Informação foi a mais genérica de todas as áreas. É patente a escolha pela opção CIÊNCIAS SOCIAIS APLICADAS/CIÊNCIA DA INFORMAÇÃO. A limitação da amostragem analisada nesta pesquisa impede de se chegar a conclusões mais detalhadas sobre este fato, porém, arrisca-se dizer que os autores, diante da precariedade de representação da árvore de conhecimento, optaram por pecar por generalidade, evitando subdivisões mais específicas.

4.3 DISCUSSÕES E SUGESTÕES

A escolha da **PL** como objeto de estudo desta pesquisa deu-se, particularmente, por duas razões: pela importância e credibilidade que o sistema conquistou ao longo do tempo, mas principalmente - e este é foco do trabalho - por se tratar de um sistema de informação caracterizado por um processo de preenchimento aberto.

Os conceitos de sistemas abertos e fechados remontam a discussões oriundas das idéias de Bertalanffy, que durante a década de 1960 criou a Teoria Geral dos Sistemas. Segundo Machado (2003), a percepção de uma abordagem sistêmica, pregava uma contínua revisão do mundo, do sistema como um todo e de cada um de seus componentes. A partir dessa visão sistêmica, Bertalanffy (1977) compreendeu que os sistemas poderiam ser

fechados ou abertos (na prática nenhum chega a ser totalmente fechado ou aberto).

Os sistemas fechados são aqueles com pouca interação com o meio ambiente que os circunda. Tais sistemas mantêm, com relação ao meio externo, poucas entradas e saídas e, por esta razão, o sistema fechado é também chamado sistema mecânico ou determinístico. Os sistemas abertos interagem mais com o meio, adaptando-se às mudanças em busca da própria sobrevivência, mantendo contínuas interações com o ambiente que o envolvem. Em sistemas abertos a vulnerabilidade decorre do baixo nível de controle da situação, enquanto que nos fechados "o estado final é inequivocamente determinado pelas condições iniciais" (BERTALANFFY, 1977, p. 64).

A Teoria Geral dos Sistemas trouxe à tona princípios que - refletidos na questão de sistemas de informação - fazem perceber o quanto qualquer sistema, natural ou cultural, é influenciado pelo nível de interação que o mesmo tem com o ambiente no qual está inserido.

A **PL** é um sistema aberto. Numa terminologia da Ciência da Informação, isso implica dizer que se trata de um sistema com um baixo nível de controle. O controle diz respeito às representações lingüísticas que são usadas no preenchimento dos currículos, excluindo-se, portanto, questões de segurança no acesso ou rastreamento do comportamento/interação dos usuários com o sistema.

O planejamento/desenvolvimento da **PL**, intencionalmente ou não, desconsiderou as vantagens proporcionadas aos sistemas pelo controle que a eles podem ser atribuídos. Em contrapartida, foi beneficiado por um grande ganho econômico ao compartilhar com a comunidade acadêmica o compromisso de alimentar um sistema que serve de apoio aos órgãos de fomento brasileiros. Reduz-se o custo de investimento mas, em compensação, perde-se consistência nas informações disponibilizadas. Princípio semelhante (apenas no aspecto de alimentação dos sistemas) ocorre com os diversos repositórios abertos – também chamados de arquivos abertos - voltados ao ambiente da C&T. E seguindo o mesmo princípio, o de tornar o usuário um

agente ativo nas representações de conteúdos – existem as folksonomias que, entretanto, não se restringem ao universo da ICT[62].

Na **PL,** a perda de consistência na RI, conforme visto nos exemplos analisados, poderia ter sido menor se fossem adotados procedimentos orientados ao controle do sistema. Tais procedimentos podem ser utilizados não somente nos sistemas de currículos, mas para todos os que utilizam termos para representação de informações. Desta forma, são apresentadas a seguir recomendações voltadas à organização da informação, que apesar de já bastante difundidas nos domínios da Ciência da Informação e de não serem inéditas, podem contribuir para a concepção e funcionamento de sistemas eletrônicos de informação.

O controle de vocabulário inicia com procedimentos que Smit e Kobashi (2003) denominaram de "micro" e que servem ao controle nos termos ou expressões em arquivos, tais como: a) Normalização gramatical, b) Opções de grafia, c) Controle de sinonímia e d) Controle de homonímia. O procedimento "macro" diz respeito à organização dos termos em formatos previstos em classificações ou tesauros. Cada procedimento acima enumerado será detalhado a seguir.

a) Normalização gramatical: recomenda-se a adoção da forma substantiva, masculina e singular dos termos;

É possível incluir em sistemas como a **PL** recursos similares aos utilizados em corretores ortográficos dos editores de texto (como o Microsoft Word). Isso evitaria, no mínimo, erros elementares de digitação. Mas, além de corrigir erros, o recurso seria mais proveitoso se funcionasse a partir de um vocabulário de termos criado para áreas específicas. Tal atitude seria imprescindível para possibilitar a identificação de termos adequadamente, levando-se em conta a adoção da forma no substantivo, masculina e singular dos termos. O problema atual é como proceder diante dos mais de um milhão de currículos cadastrados na **PL** e da crescente interdisciplinaridade entre áreas do conhecimento.

Uma alternativa é aplicar técnicas de mensuração de palavras para contabilizar a freqüência dos termos mais recorrentes. Identificados, os termos

[62] Questões discutidas na seção 4.2.2.

poderiam ser reformatados a partir das recomendações de normalização gramatical, ressaltando-se que essa tarefa deve contar com a participação de especialistas da área em razão de se trabalhar com linguagem bastante especializada. Com a lista de termos recomendados, o processo de substituição na base é passível de ser automatizado. Porém, uma alteração não autorizada pode ser alvo de críticas, razão pela qual seria melhor apresentar sugestões de mudança aos usuários nos itens que forem necessários.

Essas sugestões são de caráter corretivo, destinadas a reduzir inconsistências cuja adequação é relativamente fácil. Considerando a quantidade de currículos cadastrados, é de se esperar – baseando-se nos exemplos vistos – um alto índice de ajustes que favorecerão a **PL** como um SICT. Evidentemente que as correções dependerão da boa vontade dos "proprietários" de cada currículo.

b) Opções de grafia: o procedimento costuma envolver situações na qual o mesmo termo ou expressão apresenta grafias diferentes (geralmente em razão da passagem do tempo). Esta ação pode envolver três aspectos distintos de um SICT: o planejamento, a manutenção e a correção. Para o planejamento, que é uma fase anterior à inserção dos conteúdos, é importante prever mecanismos de orientação àqueles que alimentarão o sistema.

Os aspectos de manutenção e correção são inter-relacionados. A manutenção deve ser feita pelos gerentes dos sistemas de informação, que com o auxílio de especialistas das áreas, podem atualizar a lista de termos no que se refere às opções de grafia. Com a lista atualizada, recomenda-se o uso de remissivas que orientem os usuários na escolha do termo. Com as remissivas, os novos registros que porventura fossem utilizar termos em desuso serão orientados a adotar o termo preferido pelo sistema. No caso dos registros anteriores à atualização de determinados termos, será preciso estabelecer uma rotina que identifique os currículos com tais registros e que recomende ao usuário a correção necessária.

Para a **PL,** a mesma estratégia sugerida na normalização gramatical também seria válida para opções de grafia, mas o trabalho provavelmente seria maior, pois além de haver a necessidade de identificação dos termos mais

usados na base do sistema, seria preciso analisá-los conforme as áreas de conhecimento, para identificar quais opções de grafia deveriam prevalecer.

c) Controle de sinonímia

A sinonímia é uma relação de equivalência entre, ao menos, duas palavras. Para um sistema de informação interessa o quanto um termo é preferencial para ser utilizado no sistema com relação a outros termos. O ideal é representar o conceito através de um único termo e assim, permitir a combinação entre a linguagem do usuário e a do sistema. Porém, se o ideal é o uso de um termo único, o contrário (uso de vários termos) deve ser evitado por ser prejudicial ao processo de recuperação da informação. O uso de muitos termos dificulta a compatibilização entre uma estratégia de busca e as formas de representações lingüísticas na base, além de provocar uma dispersão de informações devido ao uso de vários termos para um mesmo conceito.

A título de exemplo, se a relação sinonímica entre as expressões PROCESSOS DISSIPATIVOS e PROCESSOS IRREVERSÍVEIS é intuitivamente clara para aqueles que atuam na área da Física, mas o mesmo não é verdade para a **PL**, que como qualquer sistema eletrônico de informação, requer uma rotina que estabeleça formalmente a equivalência entre os termos.

Um recurso que permita ao sistema “compreender” que dois ou mais termos diferentes têm o mesmo significado não é trivial. É requerido um grau de especialidade relativamente alto para prever relações de equivalência entre termos e/ou expressões, ou seja, é preciso dominar a respectiva área do conhecimento. Acrescenta-se que as relações devem seguir um pressuposto nocional capaz de interpretar o significado para determinado domínio ou área de conhecimento para a qual as equivalências devem e podem ser estabelecidas.

Para um sistema como a **PL,** essa é uma missão bastante penosa – ou até impossível - pois exige a capacidade de lidar com todas as áreas de conhecimento. O fato de a **PL** abarcar domínios de conhecimento de toda a C&T implica em montar esquemas de relações para cada um dos domínios. Criar um esquema único, capaz de associar universos tão diferentes do conhecimento, seria uma tarefa extremamente complexa – **quiçá impossível**,

em todo caso fadada ao insucesso - dada a multiplicidade de universos semânticos.

A necessidade de contextualização das relações é condição básica. Isso foi percebido no início deste trabalho, quando se buscou uma definição adequada para a expressão “organização da informação”. Nesse exemplo incorreu um caso de polissemia, que é o fenômeno pelo qual uma palavra ou expressão pode comportar mais de um significado dependendo do contexto de seu uso. Exemplo: Organização da informação (área da Arquitetura), Organização da informação (área da Ciência da Informação).

Outro caso importante no âmbito da C&T é a mudança de termos decorrentes da consolidação da terminologia da área: não são raros os casos de uso de um termo que, com o passar do tempo, cai em desuso ou então se transforma em um outro, que passa a vigorar na linguagem da área. Termos como “MEIO AMBIENTE”, “AIDS” ou “PORTADOR DE NECESSIDADES ESPECIAIS”, por razões diferentes, foram cunhados recentemente, fruto de processos sociais que sempre estarão presentes no ambiente da C&T.

Outro recurso necessário à **PL** é a identificação da mudança no nome do pesquisador. Tal modificação pode ocorrer, por exemplo, devido a casamento civil, de uma opção do autor em usar formas diferentes para registro do seu nome. Um recurso relativamente simples e pouco oneroso é recomendado: todo cadastro de currículo está vinculado a um número identificador único – o do CPF do pesquisador. Assim, uma rotina no sistema poderá registrar, para determinado CPF cadastrado na base de currículo, quaisquer alterações no nome feitas em SICT nacionais, a partir disso, indicar em que períodos cada forma vigorou.

Igualmente importante seria a viabilidade de interoperabilidade entre os SICT, no que se refere ao nome do pesquisador. Esta não seria uma atividade complexa (no tocante à organização da informação, desconsiderando-se questões tecnológicas).

Se a interoperabilidade entre a **PL** e a SciELO possibilitasse uma comparação automática entre os nomes de pesquisadores, certamente seriam apontadas as diferenças que foram verificados nos exemplos mostrados nas análises de currículos (p.93). Um recurso dessa natureza poderia ser ampliado para outros SICT nacionais e seria útil para orientar o pesquisador a adotar

uma única forma para seu nome. Talvez uma base única de nomes de pesquisadores fosse vital para a ICT brasileira, associando os nomes aos respectivos CPFs. Vale lembrar que as bases de dados da Coordenadoria de Aperfeiçoamento do Ensino Superior (CAPES) identificam todos os pesquisadores pelo respectivo CPF.

d) Controle de Homonímia

A homonímia é o fenômeno pelo qual diferentes entidades são designadas pela mesma palavra. Ela ocorre entre itens com significados diferentes que possuem o mesmo som e a mesma grafia (homônimos perfeitos: como “literatura” (substantivo) e “literatura” “disciplina”), ou o mesmo som (homônimos homófonos: caça (ato de caçar) e cassa(tornar sem efeito)), ou apenas a mesma grafia (homônimos homógrafos: como o verbo “seco” e o adjetivo “seco”).

Na **PL** a homonímia torna-se um problema muito mais grave em razão de dois fatores:

Primeiro, o SRI não é capaz de diferenciar as mais simples relações sintáticas, ou seja, numa procura pelo termo porta são recuperáveis todos os currículos nos quais a palavra PORTA estiver presente e ainda nos currículos com a palavra PORTA como radical, exemplo: PORTA-enxertos, comPORTAmento, imPORTAção.

Segundo, além do problema com os radicais, o sistema não evita palavras irrelevantes para os processos de recuperação da informação, as chamadas *STOPWORDS,* geralmente compostas de preposições, artigos ou conjunções[63].Em tal ocorrência, se for feita uma busca por PARÁ (estado brasileiro) o resultado considerará todo o currículo que contenha a preposição PARA.

Inicialmente, é urgente a necessidade de filtragem das *Stopwords* no sistema de recuperação da **PL**. Nesse caso, é preciso criar uma lista de termos indesejáveis (excetuados os casos em que os mesmos compõem sintagmas) e tais termos devem ser desconsiderados pelo sistema quando o mesmo gerar a lista de índices. Este é um procedimento interno, que não envolve o

[63] Em alguns casos como nos termos compostos, o uso de preposição e outras stopwords dão significado ao termo, exemplo: DOR-DE CABEÇA, CLINICA DE REPOUSO, TECNOLOGIAS DE INFORMAÇÃO, CIÊNCIA DA INFORMAÇÃO. Neste caso trata-se de sintagmas, que devem ter um tratamento diferenciado a partir de sua identificação.

preenchimento dos currículos, pois não há como sugerir que os usuários evitem preposições, artigos, conjunções, advérbios e outras palavras comumente consideradas *stopwords*, pois se tornaria inviável o preenchimento de campos que utilizam a linguagem natural, tais como o campo TÍTULO.

Para um melhor entendimento dos sintagmas cita-se Cabré (1993), que classifica termos a partir de sua *forma*, *função*, *significado* e *procedência*. Para fins de discussão de questões relacionadas à homonímia, interessa apenas a primeira categoria. Quanto à forma, os termos são classificados a partir do número de morfemas, podendo ser simples ou complexos. De acordo com os tipos de morfemas, os termos complexos, subdividem-se em termos derivados e termos compostos. Os termos derivados são formados pela junção de afixos a uma base lexical.

Os termos compostos (também denominados de sintagmas) são freqüentes em domínios especializados e podem ser formados pela soma de dois termos ou, até mesmo, por uma construção sintagmática mais complexa. Assim, os termos compostos são formados por palavras ou por radicais que pertencem a classes de palavras diversas. A seguir, são enumerados alguns sintagmas que foram extraídos a partir das palavras-chave dos exemplos analisados na SciELO: Coelho Doméstico, Letramento Digital, Mecânica Estatística, Membrana de Barreira, Campos Cristalinos, Saúde Coletiva, Impacto Bibliográfico, Mecanismos de Busca, Cenários Futuros, Tempos de Relaxação.

e) Organização dos termos

Ações sistematizadoras, que exigiriam maior esforço, cabem nas atividades que Smit e Kobashi (2003) chamaram de procedimentos MACRO.

As discussões referentes aos procedimentos micro abrangeram sugestões para adoção de um maior controle de termos, atividade compreendida como controle de vocabulário. Porém, se os termos controlados não forem ordenados de acordo com um critério, o vocabulário controlado será uma mera lista de termos, cujo significado se restringirá aos próprios termos. Ordenar os termos introduz no controle de vocabulário uma forma de organização dentro de um sistema significante.

Uma simples lista não apresenta significados, ou previsões sobre um domínio específico, tampouco o ponto de vista adotado e nem o nível de especificidade no qual a documentação foi tratada. É recomendável

> que os termos, uma vez submetidos ao controle de vocabulário, sejam ordenados, organizados ou categorizados. A categorização gera significado ao introduzir os termos num sistema significante" (SMIT e KOBASHI, 2003, p.34).

Entende-se que caberia ao planejamento de um SICT da dimensão da **PL** a elaboração de contextos de organização da informação segmentados por áreas. Na prática seria elaborar estruturas significantes de termos para domínios específicos de conhecimento. Assim, para pesquisadores da área da Ciência da Informação, deveria haver um instrumento dotado de termos da própria área que os auxiliasse no preenchimento dos campos.

Os desenvolvedores da **PL**, a partir da Tabela de Áreas do Conhecimento do CNPq, criaram um recurso (Figura 18) que conduz o usuário na escolha, dentro de um plano classificatório, de área(s) do conhecimento referentes às suas produções bibliográficas, técnicas, ou artísticas/culturais.

Figura 18 – Tabela de Áreas do Conhecimento do CNPq

Essa classificação das Áreas do Conhecimento, usada pela **PL,** segue uma estrutura arborescente similar a um plano de classificação, que é um tipo de vocabulário controlado. Ambos – a classificação da **PL** e um plano de

classificação – têm por base o princípio da hierarquia que oferece como vantagem o fato de, ao ordenar as atividades hierarquicamente, possibilitar uma visão do conjunto e de como essas se distribuem. A desvantagem está na necessidade de se ampliar o universo de escopo com níveis mais complexos.

Na figura 18, é nítida a subdivisão hierárquica em quatro níveis, decrescentes do nível genérico ao mais específico. A **PL** adota recurso semelhante para determinar os Setores de Aplicação das produções dos pesquisadores. Trata-se de uma classificação menos estruturada, com somente dois níveis na hierarquia. Um exemplo (Figura 19) de um Setor de Aplicação é “Desenvolvimento de programas (software) e prestação de serviços em informática”, com as seguintes subdivisões:

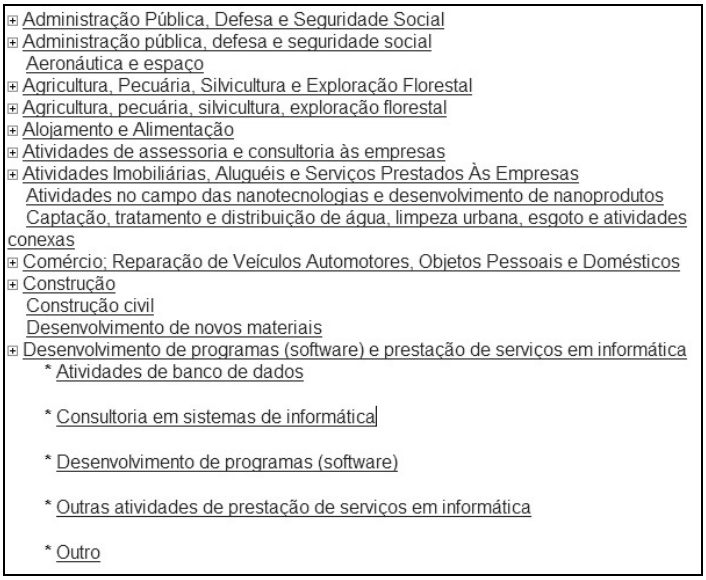

Figura 19 – Exemplo de parte da Tabela de Setores de Aplicação

Do ponto de vista da organização da informação, esses dois recursos da **PL** partem do pressuposto de que, tanto as Áreas de Conhecimento quanto os Setores de Aplicação e suas respectivas subdivisões, organizam-se em classes auto-excludentes. Tal pressuposto, face ao disposto pela Teoria da Classificação desenvolvida na área da Biblioteconomia, é correto. Ressalta-se que o sistema não impede que o usuário cadastre mais de uma opção, ou acrescente informações no nível mais específico. A inclusão de novas opções pelo usuário, no entanto, abre a possibilidade da inclusão de sinônimos ou uma classe que não seja auto-excludente em relação aos termos já previstos pelo sistema.

A Figura 20 mostra que é possível, por exemplo, cadastrar uma nova sub-área do conhecimento (Profissional da Informação), embora ela integre, na opinião de diferentes autores, a Biblioteconomia, Arquivologia e a Museologia.

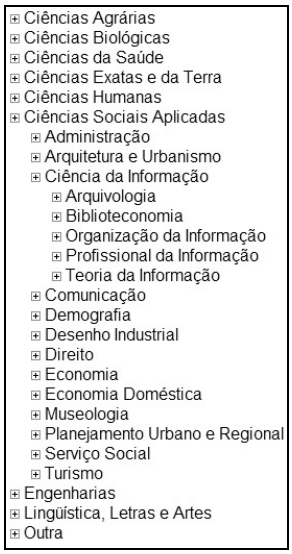

Figura 20 – Exemplo de cadastramento de nova sub-área

Numa primeira visão, é possível entender que, para um sistema voltado a um contexto informacional tão amplo como a **PL**, seriam necessários (mesmo que somente para indicar as Áreas de Conhecimento e os Setores de Aplicação) níveis de especificidade mais aprofundados ou talvez mais categorias em cada nível. Entretanto é justificável a opção generalista e reducionista dos projetistas do sistema: os dados coletados nestes campos são utilizados para fins de produção de indicadores, e como se sabe, quanto mais dispersos e fragmentados, menor será a possibilidade de se estabelecerem agrupamentos homogêneos, suficientemente capazes de demonstrar algum comportamento das sociedades científica ou tecnológica.

Contudo, seria interessante que o sistema contemplasse um número maior de níveis hierárquicos. A partir desta maior especificidade da informação seria possível, por exemplo, perceber como os pesquisadores subdividem a área de conhecimento e, ainda, definir níveis hierárquicos que poderiam ser adotados para a produção de indicadores.

Por outro lado, existe também a possibilidade do preenchimento com termos generalizantes ser percebida como desestimulante em razão dos pesquisadores encontrarem dificuldades para relacionar suas produções com termos generalistas oferecidos pelo sistema. Um sinal dessa situação foi

percebido nos exemplos examinados neste estudo. Verificou-se que o campo Setores de Atividades é pouco preenchido e o de Áreas do Conhecimento apresenta, freqüentemente, diferentes escolhas entre os autores de um mesmo artigo (lembrando que autores diferentes de um mesmo artigo podem ter visões distintas). Para conclusões mais precisas, uma análise mais criteriosa – orientada a procedimentos estatísticos de amostragem - seria necessária: fica aqui o registro da sugestão para futuras pesquisas.

Imagina-se que para o estado atual da base de currículos da **PL**, uma possível ação seria analisar se os campos específicos ÁREAS DE CONHECIMENTO E SETORES DE APLICAÇÃO conseguem indicar de forma consistente comportamentos do contexto da C&T brasileira[64]. Para tal, são necessários estudos que não podem ser restritos à mera quantificação da produção científica e tecnológica, pois os números só produzirão significado se interpretados a partir do conjunto das políticas relacionadas ao universo da C&T.

Se uma análise da base de currículos da **PL** conseguir responder às necessidades para as quais o sistema foi desenvolvido, entende-se que uma mudança não é prioritária, caso contrário uma avaliação da classificação das Áreas de Conhecimento e Setores de Aplicação será importante, visando uma provável reformulação. Tal ação é indicada considerando que o problema aumentará numa razão proporcional à inclusão de mais pesquisadores e também ao aumento da produção científica, técnica e artística nacional.

Caso se comprove a necessidade de reformular a classificação, sugere-se observar o processo de construção de um plano de classificação que incorpore o controle de vocabulário. Tal tarefa exige a composição de equipes formadas por atores especialistas nas respectivas áreas de conhecimento e também atores capacitados para elaborar ferramentas orientadas à organização da informação. A sugestão de procedimentos segue os passos descritos por Smit e Kobashi (2003):

1 - levantamento das listas livres (no caso da **PL**, uma para Áreas de Conhecimento e outra para Setores de Atividades);

[64] Sabe-se que uma nova Tabela de Áreas do Conhecimento foi elaborada recentemente, mas não consta que a mesma tenha sido aprovada ou tornada pública.

2 - análise crítica[65], se necessário, dos termos incluídos nas listas (verificar sinonímias, consistência em termos de normalização gramatical, opções de grafia e solução dada aos termos compostos);
3 - elaboração de listas alfabéticas consistentes de áreas e setores, desdobrada em suas respectivas especificidades, caso for preciso, e acrescida das remissivas que se fizerem necessárias;
4 - categorização, em maior ou menor grau, dos itens constantes da lista. Nomeação das categorias maiores, novamente incorporando a preocupação com o controle de vocabulário;
5 - análise das nomeações que podem gerar leituras diferentes e elaboração de notas de escopo ou notas de uso[66]. Tanto as notas de escopo como as de uso serviriam como recursos de orientação para o preenchimento da **PL**. A elaboração destas (sobretudo as de escopo) requer a participação efetiva dos especialistas da área para atribuir, a partir do domínio de conhecimento específico, o conceito próprio ao termo. Os especialistas da área da informação seriam necessários para orientar sobre a importância, a função e, principalmente, a elaboração das referidas notas;
6 - submissão das listas (modalidade, categorizada e alfabética) a testes, avaliação do resultado dos testes, incorporação de ajustes e efetiva implantação do plano de classificação.

Outra possibilidade de organização das Áreas de Conhecimento e Setores de Aplicação é proposta pelo tesauro. O tesauro e o plano de classificação são instrumentos para organização da informação que incluem o controle terminológico em graus diferenciados, e são utilizados em sistemas de informação visando traduzir a linguagem dos documentos, dos indexadores e dos pesquisadores numa linguagem controlada, para uso na indexação e recuperação de informações.

Os tesauros apresentam maior flexibilidade na sua elaboração, pois não partem do princípio de uma única hierarquia para organizar os termos. No entanto, para os propósitos desta discussão, reforçamos a necessidade da

---

[65] A análise dos termos deve considerar o aspecto temático (averiguação feita por especialistas da área para analisar se é ou não cabível a inclusão do termo no domínio terminológico da área) e técnico (verificação por especialista da área da informação para analisar se a forma de registro do termo se adequa às recomendações visando o controle de vocabulário).
[66] Segundo Smit e Kobashi (2003, p.31-2) "as notas de escopo têm por finalidade explicitar a amplitude ou o entendimento atribuído ao conceito. [...]. As notas de uso, pouco utilizadas na prática, não se propõem a explicitar o conceito mas a explicitar recomendações práticas que devem nortear o uso do termo."

categorização dos termos, quer seja no contexto de um plano de classificação ou de um tesauro.

Um aspecto mais recente sobre os tesauros relaciona-os às ontologias. Uma discussão sobre essa questão foi feita na seção 3.4.1, mas, considera-se importante retomar o debate neste ponto do trabalho. De fato, há entre os tesauros e as ontologias algo em comum: ambos configuram um sistema de conceitos, porém, conforme estudo de Moreira, Alvarenga e Oliveira (2004), os tesauros servem de instrumento de registro e controle terminológico, para uso humano, ao passo que as ontologias objetivam o registro do conhecimento para inferências computacionais.

A posição das autoras é verificada na ontologia estabelecida para a **PL** através do *CONSCIENTIAS*. A Ontologia da **PL** é uma detalhada estrutura das partes que compõem o currículo, representando associações e níveis de subordinação/equivalência entre essas partes. Nesta estrutura há classes e categorias, com seus respectivos atributos, ou seja, na Classe Produção uma parte da estrutura assemelha-se à representação a seguir (Figura 21):

PRODUÇÃO BIBLIOGRÁFICA
ARTIGO PUBLICADO
DADOS BASICOS DO ARTIGO (ATRIBUTOS: IDIOMA, MEIO DE DIVULGAÇÃO, ANO, PAIS DE PUBLICAÇÃO, TC.)
DETALHAMENTO DO ARTIGO
AUTORES
ELEMENTOS COMUNS A ITEM PRODUÇÃO
PALAVRAS-CHAVE (ATRIBUTOS: PALAVRA-CHAVE 1, PALAVRA-CHAVE 2..)
ÁREAS DO CONHECIMENTO (ÁREA DO CONHECIMENTO 1, ÁREA DO CONHECIMENTO 2)
SETORES DE ATIVIDADE (SETOR DE ATIVIDADE 1, SETOR DE ATIVIDADE 2)

Figura 21 – Parte da ontologia da Plataforma Lattes

A Figura 21 demonstra que a relação se dá entre classes e não entre conceitos, ou seja, não existe uma relação nocional, pois não há significado semântico sob o ponto de vista humano. Existe, sim, um sentido dado ao currículo entre as partes que o compõem. Na prática, essas relações, por si só, exprimem somente esquemas de relações genéricas e relações partitivas.

Essas relações serão, de fato, utilizadas quando o sistema de informação que adotar uma determinada ontologia iniciar a inserção de dados em sua base. Na **PL,** a ontologia tem uso efetivo a partir dos currículos cadastrados e é a partir deles que podem ser executadas as inferências computacionais. O processamento automático das inferências será, então,

responsável por classificar conceitos dentro de uma hierarquia e ainda verificar se determinadas instâncias pertencem a determinadas classes.
Exemplo: Se existe registrada como título de um livro a frase: *CONHECIMENTO PÚBLICO*, a ontologia já terá previsto que aquele campo específico sempre pertencerá à instância de uma produção bibliográfica específica (livro) e aquele campo sempre "significará" o título deste livro.

A partir da ontologia, são elaborados modelos lógicos para verificar inferências, conforme as mais diversas finalidades. Um exemplo: deseja-se saber se nos últimos três anos os bolsistas de produtividade da área de Psicologia convergiram suas produções bibliográficas para itens mais importantes, segundo critérios do CNPq. Suponha-se que publicar em periódicos da Qualis[67] da CAPES seja um desses itens importantes. Seria possível, através dos currículos dos pesquisadores, associar as classes referentes à produção bibliográfica a um sistema externo que, neste exemplo, é a Base Qualis da CAPES. O modelo buscará relacionar as classes e categorias atinentes às instâncias concernentes à produção bibliográfica de artigos. Para que isso seja possível, as ontologias da **PL** e da Base Qualis deveriam ser compatíveis.

A compatibilização entre sistemas é feita automaticamente e poderá se repetir quantas vezes for desejada. E, se a inferência for consistente, assim se manterá desde que a ontologia não seja alterada. Ressalta-se que as ontologias, caso compartilhadas, permitem que, se um sistema utilizar a mesma ontologia da **PL**, esse poderá processar os mesmos modelos lógicos e inferências que porventura forem utilizados para a **PL**.

Uma das deficiências das ontologias, se comparadas ao tesauros, está na dificuldade para criar relações semânticas direcionadas a representações de conceitos. Na **PL**, por exemplo, a ontologia não altera a representação feita pelo autor através de palavras-chave de um artigo por ele publicado. A ontologia poderá fazer uso do que foi preenchido, mas não ajudará a preencher. Então ela – a ontologia – funciona na **PL** como uma meta-estrutura que pode viabilizar relações semânticas, mas não as realiza. No exemplo a

[67] Segundo http://www.capes.gov.br/avaliacao/webqualis.html "Qualis é uma lista de veículos utilizados para a divulgação da produção intelectual dos programas de pós-graduação stricto sensu (mestrado e doutorado), classificados quanto ao âmbito de circulação (Local, Nacional, Internacional) e à qualidade (A, B, C), por área de avaliação. A Capes utiliza o Qualis para fundamentar o processo de avaliação do Sistema Nacional de Pós-Graduação".

seguir (Figura 22) são apresentadas, de forma muito simplificada, duas estruturas fictícias de sistemas que compartilham de uma ontologia voltada à produção na C&T.

SISTEMA A	SISTEMA B
ARTIGOS **Periódico:** Transinformação **Ano:** 2002 **Autor:** MONTEIRO, S.M. **Título:** Uso de vocabulários controlados na Web. **Palavras-chave**: Vocabulários controlados; Web. **Área**: Ciência da Informação	ARTIGOS **Periódico:** Ciência da Informação **Ano:** 2002 **Autor:** SILVA, R. H. **Título:** Os tesauros como ferramentas para organizar conteúdos na Internet. **Palavras-chave:** Tesauros; Internet. **Área**: Ciência da Informação

Figura 22 – Exemplo de duas estruturas fictícias de ontologias

Um modelo lógico seria capaz de inferir que nos dois sistemas há produções bibliográficas da área da Ciência da Informação, da classe artigo, publicados no ano de 2002. Do ponto de vista semântico, os dois artigos assemelham-se por discutirem temáticas semelhantes, mas este entendimento exigiria uma compreensão de conceitos a partir do domínio da área da Ciência da Informação, e a ontologia não contempla essa interpretação de significados humanos.

Observou-se que muitas das inconsistências verificadas nos exemplos analisados ocorreram por conta de problemas de preenchimento. Nesse caso, a adoção de tesauros em conjunto com a ontologia da **PL** contribuiria, por exemplo, para que o preenchimento de palavras-chave fosse realizado com o auxílio de um tesauro para cada área de conhecimento, com termos e relações próprios ao seu universo de significados, resultante de um maior controle terminológico.

Considera-se que o uso híbrido – tesauro e ontologia – seria de grande valia para os SICTs proporcionando, entre outros benefícios, a possibilidade de busca orientada através de disponibilização de tesauro na etapa de definição dos termos para busca e inferências entre SICTs distintos, desde que compartilhem de ontologias comuns. O uso híbrido para a organização da informação não é uma sugestão recente, já que no artigo de E.W. Dias (2001) o autor recomenda o uso combinado de instrumentos desenvolvidos especificamente para o contexto digital com recursos que já eram utilizados antes da adoção das tecnologias eletrônicas para fins de organização da informação.

Sistematizando as discussões, observou-se que a padronização dos vocabulários, segmentando-os por áreas específicas de conhecimento, possibilitaria a identificação mais adequada de termos adotando-se a forma no substantivo, masculino e singular dos termos. Para a normalização gramatical da **PL** é necessário identificar e listar termos mais usados pelos usuários do sistema, em seguida analisá-los segundo as áreas de conhecimento, para então identificar quais opções de grafia prevaleceriam.

Idealmente, a normalização gramatical em repositórios abertos deve prever ações importantes na etapa inicial do desenvolvimento do sistema. Investir na elaboração prévia de uma lista de termos especializados é um bom caminho, porque, além de tudo, é importante que a SICT desenvolva atividades compartilhadas para o controle de vocabulários.

Outra situação identificada que requer atenção é a mudança nos nomes próprios e/ou a utilização de mais de uma forma para esses nomes. É imprescindível controlar essas diferenças e um caminho relativamente simples já foi recomendado: um rigoroso controle dos nomes através do CPF do pesquisador. Em caso de alteração o sistema identificará a mudança e tomará as devidas providências.

Para o planejamento de um SICT devem ser previstos procedimentos adequados conforme a finalidade do sistema, e isso precisa ser estipulado na fase inicial da idealização do projeto. Assim, além da normalização gramatical anteriormente referida, vale investir na elaboração prévia de uma lista de termos especializados: para tanto a formação de equipes para cada área de conhecimento é imprescindível, pois estas seriam as responsáveis pela elaboração da lista de termos. As equipes pressupõem especialidade nas respectivas temáticas.

É igualmente importante para os SICT´s desenvolverem atividades compartilhadas, em que seria fundamental a participação de uma instituição como o IBICT para formular uma política nacional que estimule convênios entre sistemas. No que tange ao controle de vocabulários, percebe-se que estão sendo criados diversos caminhos para a organização da informação. Na SciELO, por exemplo, existe uma lista controlada de termos que foram gerados a partir das palavras-chave dos artigos dos periódicos disponibilizados nesta biblioteca digital. Outro exemplo é o da BIREME que criou uma interessante

ferramenta denominada DEC´S (Termos em Ciências da Saúde) que, como o próprio nome sugere, é um conjunto de termos na área de saúde, que foi formulado a partir do MeSH (Medical Subject Headings).

Uma das discussões atuais sobre os conteúdos da Internet volta-se para utilização compartilhada de recursos entre sistemas disponibilizados na grande rede, discussão essa relacionada com os princípios de Web Semântica. Desta maneira, a utilização de vocabulários controlados por parte dos SICT nacionais deve ser estimulada entre os atores que gerenciam tais sistemas, para averiguarem a viabilidade de implantações conjuntas. Reconhece-se que a compatibilização semântica entre os sistemas não é simples, seria uma quase remodelagem do antigo sonho do controle bibliográfico universal (em nível nacional). Mas a adoção, mesmo que reduzida ou simplificada de vocabulários controlados em sistemas com pouco ou nenhum controle, pode ser benéfica no que diz respeito à organização da informação.

Uma questão mais delicada na **PL** são as polissemias, homonímias e sinonímias, pois o referido sistema abarca a totalidade de áreas de conhecimento da C&T, o que implica em esquemas de relações para cada domínio dada a impossibilidade de sistema nocional único.

Entende-se que caberia ao planejamento de um SICT, da dimensão da **PL**, a criação de contextos segmentados de organização da informação. Na prática, seria necessário elaborar estruturas significantes de termos para domínios específicos de conhecimento. O problema maior reside em pensar nesse aspecto como uma ação corretiva, quando idealmente haveria de ser uma atividade da fase de idealização/planejamento do sistema.

De qualquer forma, propõe-se seguir um caminho similar à normalização gramatical: aferir numericamente os termos mais usados, compor equipes especialistas por áreas de conhecimento e montar um vocabulário controlado que, num primeiro instante, buscará corrigir as inconsistências atuais do sistema e, posteriormente, proporcionará um controle maior no preenchimento dos currículos. Estamos cientes, no entanto, que este procedimento, embora indispensável, em nossa opinião não resolverá de forma duradoura a questão, pois a linguagem das várias áreas do conhecimento é dinâmica e as abordagens interdisciplinares talvez não sejam reconhecidas enquanto tal e, consequentemente, acabem sendo inseridas no vocabulário de outras

disciplinas. De toda forma, para outros SICTs que venham a ser criados numa concepção de sistemas abertos, necessário se faz considerar questões mínimas de controle.

Por fim, sugere-se o uso combinado de um tesauro e da ontologia já estabelecida para a **PL**. Cada um desses recursos tem funções próprias, que não se anulam e tampouco se sobrepõem, mas se complementam. A ontologia não é capaz de favorecer as representações conceituais em forma de palavras, coisa que o tesauro é reconhecidamente capaz de fazer, podendo ser usado tanto pelos usuários que preenchem o sistema, como pelos que buscam informações nele. Ao tesauro não cabe a tarefa de criar uma estrutura de relações que possam produzir inferências lógicas entre suas partes, esse é o papel da ontologia. A ontologia ainda permite o compartilhamento de suas estruturas com outros sistemas e espera-se que os outros SICTs nacionais levem em consideração tal função da ontologia.

5 CONCLUSÃO

A problemática decorrente da natureza aberta dos SICTs nacionais foi a motivação deste estudo. E para observá-la de forma mais sistemática é que se optou por analisar a **PL**, que atualmente é um dos sistemas abertos mais utilizados pela comunidade científica e tecnológica brasileira. O expressivo número de usuários cadastrados demonstra sua amplitude e justifica a análise realizada.

Tal motivação conduziu ao objetivo maior desta pesquisa que foi discutir, avaliar e propor sugestões à organização da ICT brasileira em meio eletrônico caracterizada pela livre inserção de dados nos sistemas, tomando por exemplo a **PL**. Para alcançar esse objetivo, buscou-se: traçar um retrospecto histórico da ICT brasileira, desenvolver estudo exploratório na **PL** e relacionar os procedimentos de organização da informação utilizados pela **PL** com recursos tradicionalmente utilizados para o tratamento da informação, como os vocabulários controlados, a fim de propor melhorias

Não foi possível desenvolver análises a partir de amostragens estatísticas, mas a exposição de exemplos foi suficiente para evidenciar que há falhas oriundas do preenchimento dos currículos e da concepção do sistema. Falhas estas que desfavorecem tanto a utilização dos dados para formulação de indicadores de C&T, como prejudicam o processo de recuperação da informação quando se trata de, através da busca, identificar especialistas em determinada área ou então ter um espelho da trajetória de determinado especialista.

Apesar de indesejáveis, os erros de digitação são passíveis de soluções mais simples e, por muitas vezes, automáticas. Entretanto, as inconsistências percebidas a partir de uma visão fundamentada na organização da informação são mais críticas. Percebeu-se na concepção da **PL** o descuido em processos amplamente recomendados para a organização de acervos não-eletrônicos. Entende-se que os acervos eletrônicos possuem características inerentes ao seu formato que tornam irrelevantes certos cuidados como, por exemplo, a organização física. Mas os cuidados com elementos comuns entre meios eletrônicos e analógicos - como as representações descritivas e temáticas – devem se preservados.

Muitos problemas observados na **PL** resultaram da utilização de representações lingüísticas desaconselhadas pela Ciência da Informação, como a linguagem natural ou uso do plural. Esse uso indiscriminado de termos deve-se à falta de controle na entrada de dados no sistema, e para sistemas de informação a diversificação lingüística é algo comprometedor. Assim, sob o ponto de vista da organização da informação, visando a recuperação da informação, o uso de linguagens documentárias não é a solução para todas as questões informacionais, mas é um mecanismo reconhecidamente útil para os processos de organização e distribuição da informação, principalmente da informação especializada produzida e usada nos setores acadêmicos e técnicos.

Além da função organizadora, as linguagens documentárias estão aptas a potencializar um recurso tecnológico em expansão: a interoperabilidade (ou enlaces) entre sistemas. Tais linguagens (documentárias) são adotadas no compartilhamento de conjuntos significantes entre sistemas de informação. Já a interoperabilidade lida com o compartilhamento de padrões comunicativos entre sistemas ou padrões descritivos de documentos. Conseguir compatibilizar aspectos descritivos, temáticos e tecnológicos entre sistemas de informação é algo que os defensores da web semântica têm defendido.

A respeito da web semântica, acredita-se que a combinação entre novos instrumentos de tratamento da informação (como as ontologias e as linguagens de marcação) com outros mais tradicionais (como os vocabulários controlados) resultaria em sistemas mais consistentes e compartilháveis, tanto tecnologicamente quanto semanticamente (no sentido da compreensão humana).

Essa sugestão de uso híbrido de instrumentos não constitui uma defesa dos tradicionais recursos para tratamento da informação. É uma constatação de que certas inconsistências (como as verificadas na **PL**) dos sistemas de informação poderiam ser, no mínimo, amenizadas se determinados instrumentos e procedimentos fossem adotados. Mais oportuno que reinventar, seria repensar as ferramentas que outrora foram adotadas na organização da informação; o aperfeiçoamento delas, combinado com o desenvolvimento de novos ferramentais, possivelmente proporcionará novas e mais consistentes formas de organização da informação, visando sua recuperação.

Um exemplo prático são as folksonomias, vistas por alguns autores como uma revolução na forma de classificar conteúdos na web. Eufemismos à parte, as folksonomias, de fato, representam uma nova forma de tratar a informação, porém a literatura científica já identificou deficiências nesse instrumento e recomendações já foram feitas, muitas delas pautadas no uso de recursos conhecidos pela Ciência da Informação como os vocabulários controlados.

Como síntese das considerações a respeito da **PL**, ressalta-se que:

- o currículo gerado pelo sistema é demasiadamente longo, proporcionando uma leitura confusa e descontextualizada. Numa visão arquivística, isto o torna inadequado como formato de um documento que deveria espelhar a trajetória do pesquisador. Ainda numa visão arquivística, entende-se que o preenchimento do currículo promove a descontextualização de atividades que, originalmente, aconteceram a partir de um núcleo comum de ação. Por outro lado, é reconhecidamente positivo o fato de cientistas terem se habituado a registrarem seu histórico acadêmico, tornando possível a criação de um grande acervo de currículos de cientistas brasileiros;
- considera-se necessário inserir mecanismos de controle na forma de preenchimento da **PL**, pois o aumento da comunidade científica brasileira e o respectivo crescimento da produção desta comunidade evidenciam um aumento proporcional de inconsistências. Recursos de normalização gramatical e/ou orientações interativas que direcionem os usuários no preenchimento do currículo podem trazer benefícios a curto prazo por um custo baixo. Considera-se também inconcebível que um sistema da dimensão da **PL** mantenha em sua base de currículos falhas elementares - como erros de digitação. Assim, ações corretivas são urgentes;
- a **PL** resultou da integração de outras bases de currículos de instituições da área de C&T, já que sua concepção visava unificar informações dispersas para fins de fomento à pesquisa. Historicamente, os sistemas (e as políticas) de informação no Brasil foram descontinuados, e tornou-se corriqueiro criar novas soluções e negligenciar antigos problemas. A **PL** apresenta fragilidades quanto à organização da informação,

redundando em problemas na RI. Não há como aferir se são problemas oriundos das antigas bases incorporadas pela **PL**, pois os antigos sistemas já foram desativados. Interessa saber se serão buscadas soluções para os problemas atuais, ou então, se será aguardada uma nova solução, interrompendo um processo para iniciar um outro, deixando para trás os problemas do passado;

- corrigir as inconsistências atuais da **PL** é bem mais coerente do que aguardar a futura criação de um novo sistema capaz de solucionar as deficiências. Para tanto, é preciso rediscutir a concepção da **P**L, pensando-a não mais como uma solução integradora de bases e sim como um sistema voltado à gestão e à política de C&T. Desta forma, o planejamento deve antever quais e como as formas de preenchimento dos currículos podem servir para a geração de indicadores e a recuperação da informação.

Apesar da **PL** ter sido o objeto de estudo desta pesquisa, a problemática investigada é mais ampla, pois trata dos sistemas abertos de informação (sobretudo os de ICT), caracterizados pelo pouco ou inexistente controle na alimentação de suas bases. Sobre a referida problemática, inferiu-se que:

- somente tornar acessível a produção científica não favorece o conjunto maior da comunicação científica. A comunicação científica não é um meio, porém um processo composto por produtores, usuários e recursos que regem esse conjunto. É preciso ter clareza quanto à função de cada novo recurso informacional que será disponibilizado para a comunidade, o que requer uma definição de suas finalidades na fase de planejamento. Se a função prevista para um determinado sistema for o armazenamento de arquivos eletrônicos, aspectos de organização são secundários. Porém, se houver a expectativa de que seja um SICT dotado de recursos de recuperação da informação e/ou sirva como fonte para elaborar indicadores de C&T, é imprescindível conhecer as diretrizes necessárias próprias à organização da informação para os devidos fins;

- o uso de linguagens documentárias e conseqüente adoção de vocabulários controlados são criticados devido ao custo no processo de organização da informação. Porém, apesar da desvantagem do custo – que é real – o controle de vocabulário permite alcançar maior consistência e confiabilidade na informação tornada pública e disponível. Diante da influência da racionalidade econômica na formação dos estoques de ICT, ressalta-se que, enquanto recursos mais eficientes (e mais consistentes) não forem desenvolvidos, a referida racionalidade precisa ser refletida. No estado atual dos SICTs, não cabem mais escolhas excludentes, ou seja, adotar um controle rígido ou permitir demasiada liberdade ao sistema.

Por fim, a partir desta pesquisa, novos estudos podem contribuir para a continuidade das discussões. Entre algumas possibilidades, são sugeridos quatro caminhos promissores:

- estudos na **PL** que avaliem a consistência dos dados para fins de recuperação e análises bibliométricas. Sugere-se um recorte por áreas do conhecimento, o que poderá indicar se os currículos de determinadas áreas do conhecimento encontram-se em situação mais crítica que outras (quanto à consistência dos dados). Isso permitirá definir prioridades nas ações de melhoria do sistema;
- investigações da viabilidade de compatibilização/ compartilhamento de recursos entre sistemas abertos e sistemas fechados (controlados). Sugere-se um confronto entre os recursos (tesauros, lista de descritores, padrões de representação descritiva/temática, etc) de sistemas abertos (como a própria **PL**) com os de sistemas fechados (SciELO, Biblioteca Digital de Teses e Dissertações, LILACS da BIREME). Um estudo dessa dimensão comportaria subprojetos com atividades distintas, porém inter-relacionadas. Como resultados, poderiam surgir propostas para a ICT brasileira como: ontologias comuns aos SICTs; padrões nacionais de procedimentos para a organização de SICTs; tesauros-modelo para serem implantados como experiência em mais de um SICT ou utilizados para compatibilizar vocabulários diferentes adotados por SICTs diferentes; elaboração de listas de descritores essenciais para implantação de forma compartilhada nos SICTs;

- avaliação das políticas nacionais relacionadas à gestão dos SICTs. O objetivo será identificar se há convergência entre as ações que estão previstas para os sistemas nacionais de informação. Isto possibilitará, por exemplo, verificar se há previsão de compatibilização entre os padrões de metadados das Bases de Dados da Embrapa com os da **PL** e/ou relacionar as ontologias da **PL** com uma que sirva a Biblioteca Digital de Teses e Dissertações do IBICT;
- tendo em vista a necessidade de tornar a ICT brasileira visível internacionalmente, seria importante analisar a viabilidade de traduzir a PL para a língua inglesa ou, então, em cada registro, prever a possibilidade de inclusão do título da publicação em inglês e das respectivas palavras-chave. Ou seja, independentemente da língua original, haveria campos de título e palavras-chave em inglês. Neste caso, a língua original do texto ficaria sempre visível para o usuário.

6 REFERÊNCIAS

AGUIAR, A. C. Informação e atividades de desenvolvimento científico, tecnológico e industrial: tipologia proposta com base em análise funcional. **Ciência da Informação**, Brasília, v.20, n.1, p.7-15, jan./ jun. 1991.

ALMEIDA, M. B. e BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, Brasília, v.32, n.3, p.7-20. set./dez. 2003.

ALVARENGA, L. (2001). A teoria do conceito revisitada em conexão com ontologias e metadados no contexto das bibliotecas tradicionais e digitais. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v.2, n.6, art. 05, dez. 2001. Disponível em: < http://www.dgzero.org/dez01/F_I_art.htm >. Acesso em: 31 jul. 2002.

_____ .(2003). Representação do conhecimento na perspectiva da Ciência da Informação em tempo e espaço digitais. **Encontros Bibli**: Revista Eletr. de Biblioteconomia e Ci. Inf., Florianópolis, n. 15, p.1-23, jan./jun. 2003. Disponível em: <http://www.encontros-bibli.ufsc.br/Edicao_15/sumario_15.htm>. Acesso em: 10 jun. 2006.

_____ .(2006). Organização da informação nas bibliotecas digitais. In: NAVAES, M.M.L; KURAMOTO, H. **Organização da informação:** princípios e tendências. Briquet de Lemos/Livros: Brasília: 2006. cap. 6, p.76-98.

AMARAL, S. A. do. Serviços bibliotecários e desenvolvimento social: um desafio profissional. **Ciência da Informação**, Brasília, v. 24, n. 2, p.221-227, maio/ago. 1995.

ANDERSON, J.D.; PÉREZ-CARBALLO, J. The nature of indexing: how machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. **Information Processing & Management**, v.37, n.2, p.231-254, mar. 2001.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: ACM Press, 1999.

BARRETO, A. de A.. (1994). A questão da informação. **São Paulo em Perspectiva**, São Paulo, v.8, n.4, p.3-8, 1994.

_____. (1998). Mudança estrutural no fluxo do conhecimento: a comunicação eletrônica. **Ciência da Informação**, Brasília, v.27, n.2, p.122-127, maio/ago. 1998.

_____ .(1999). Os destinos da Ciência da Informação: entre o cristal e a chama. **Informação & Sociedade**: Estudos, João Pessoa, v. 2, n. 9, 1999. Disponível em: <http://www.informacaoesociedade.ufpb.br/IS929914.htm>. Acesso em: 05 maio 2006.

_____ .(2000). Os agregados de informação: memórias, esquecimento e estoques de informação. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v.1, n.3, jun. 2000. Artigo 01. Disponível em: < http://www.datagramazero.org.br/jun00/F_I_art.htm >. Acesso em 5 mar. 2002.

BAX, M. P. Introdução às linguagens de marcas. **Ciência da Informação**, Brasília, v. 30, n. 1, p.32-38, jan./abr. 2001.

BELKIN, N.J. Anomalous states of knowledge as a basis for information retrieval. **Canadian Journal of Information Science**, n.5, p.133-143,1980.

BERGMAN, M. K. The Deep Web: Surfacing Hidden Value. **Journal of Electronic Publishing,** v.7, n.1, aug. 2001. Disponível em: < http://www.press.umich.edu/jep/07-01/bergman.html > . Acesso em: 31 jul. 2006.

BERNERS-LEE, T. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, n.5, may 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: 13 nov. 2006.

BERTALANFFY, L. von. **Teoria Geral dos Sistemas**. 3. ed. Petrópolis: Vozes, 1977. (Teoria de sistemas, 2).

BERTERO, C. O. **Gestão de Ciência e Tecnologia:** uma análise institucional. São Paulo: Fundação Getúlio Vargas, 1994. 19 p. (Ciência e Tecnologia no Brasil: uma nova política para um mundo global).

BIOLCHINI, J. C. de A. Semântica e cognição em bases de conhecimento: do vocabulário controlado à ontologia. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v.2, n.5, out. 2001. Disponível em: < http://www.dgzero.org/Atual/Art_02.htm >. Acesso em: 25 out. 2001.

BOCCATO, V.R.C.; FUJITA, M.S.L. Estudos de avaliação quantitativa e qualitativa de linguagens documentárias: uma síntese bibliográfica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.11, n.2, p.267-281, mai./ago.2006.

BOLAÑO, C.; KOBASHI, N.Y.; SANTOS, R. N. M. dos. A lógica econômica da edição científica certificada. **Encontros Bibli**: R.Eletr.Bibliotecon. e Ci. Infor., Florianópolis, n. especial, p.119-131, 1º sem. 2006. Disponível em: <http://www.encontros-bibli.ufsc.br/bibesp/esp_03/9_GT5_bolano.pdf>. Acesso em: 30 abr. 2006.

BRANDAU, R.; MONTEIRO, R.; BRAILE, D. M. Importância do uso correto dos descritores nos artigos científicos. **Revista Brasileira de Cirurgia Cardiovascular**, São José do Rio Preto, v. 20, n. 1, 2005. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-76382005000100004&lng=es&nrm=iso>. Acesso em: 13 Nov 2007.

BRASIL. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Ministério da Ciência e Tecnologia. **Diretório dos Grupos de Pesquisa no Brasil:** Censos 2004 - 2002-2000. Disponível em: <http://dgp.cnpq.br/censo2004/index.htm>. Acesso em: 24 jun. 2006.

BUCKLAND, M.K. What is a "document"?. **Journal of the American Society for Information Science**, v.48, n.9, p.804-809, 1997.

BUSH, V. As we may think. **The Atlantic online**. Disponível em < http://www.theatlantic.com/doc/194507/bush >. Acesso em: 26 de dez. 2002. Artigo originalmente publicado em The Atlantic Monthly, n.1, p.101-108, jul. 1945.

CABRÉ, M. T. **La terminologia:** teoría, metodología, aplicaciones. Barcelona: Antártida/Empúries, 1993.

CAFÉ, L.; LAGE, M. B. Auto-arquivamento: uma opção inovadora para a produção científica. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v. 3, n. 3, p.1-2, jun. 2002. Disponível em: <http://www.dgz.org.br/jun02/Art_04.htm>. Acesso em: 27 mar. 2006.

CAMPOS, M.L. de A. (2002). A organização de unidades de conhecimento em hiperdocumentos: o modelo conceitual como um espaço comunicacional para a realização da autoria. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 20., 2002, Fortaleza. **Anais**... Fortaleza: UFC, 2002. 1 CD-ROM.

_____. (2004). Modelização de domínios de conhecimento: uma investigação de princípios fundamentais. **Ciência da Informação**, Brasília, v. 33, n. 1, p.22-32, jan./abr. 2004.

CAPURRO, R. Perspectivas de una cultura digital en latinoamerica. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v. 2, n. 3, p.1-13, abr. 2002. Disponível em: <http://www.dgz.org.br/abr02/Art_01.htm>. Acesso em: 27 mar. 2006.

CATARINO, M. E.; BAPTISTA, A. A. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v. 8, n. 3, jun. 2007. Disponível em: < http://www.dgz.org.br/jun07/Art_04.htm>. Acesso em: 21 ago. 2007.

CHATAIGNIER, M.C.P; SILVA, M. P. da. Biblioteca digital: a experiência do INP. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 7-12, set./dez. 2001.

CINTRA, A. M. et al. **Para entender as linguagens documentárias**. 2.ed. rev. e ampl. São Paulo: Polis, 2002. (Coleção Palavra-Chave, 4).

CNI; SENAI. **Demanda por informação tecnológica pelo setor produtivo:** pesquisa 1996. Rio de Janeiro: CNI, 1996.

CONSCIENTIAS. **Ontologias.** Disponível em: <http://lmpl.cnpq.br/lmpl/index.jsp?go=ontologias.htm>. Acesso em: 01 ago. 2006.

CUNHA, M. B. da. IBICT: 51 anos. **Ciência da Informação**, Brasília, v. 34, n. 1, p.7-8, jan./abr. 2005.

DAVENPORT, L.; CRONIN, B. What does hypertext offer the information scientist?.**Journal of Information Science**, v.15, n.6, p.369-372, 1989.

DIAS, E. W. Contexto digital e tratamento da informação. **DataGramaZero – Revista de Ciência da Informação**, Rio de janeiro, v.2, n.5, art. 01, out. 2001. Disponível em: < http://www.datagramazero.org.br/out01/Art_01.htm >. Acesso em: 24 jan. 2002.

DIAS, M.M.K. **O gerenciamento de unidades de informação tecnológica sob o enfoque da gestão da qualidade**: do estudo das percepções e reações dos clientes ao desenho de novas condutas. 2001. 148 f. Tese (Doutorado em Ciências da Comunicação) – Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 2001.

DIAS, P. Hipertexto, hipermédia e media do conhecimento: representação distribuída e aprendizagens flexíveis e colaborativas na Web. **Revista Portuguesa de Educação**, Minho, v. 1, n. 13, p.141-167. 2000.

DODEBEI, V. L. D. **Tesauro**: Linguagem de representação da memória documentária. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002.

FERNEDA, E. **Recuperação de Informação:** Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003. 147 f. Tese (Doutorado em Ciências da Comunicação) - Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 2003.

FONSECA, E. N. da. **Problemas de comunicação da informação científica**. São Paulo: Thesaurus, 1973. 140 p.

FUJINO, A. Política de Informação e a Hélice Tripla: Reflexões sobre Serviços de Informação no Contexto da Cooperação U-E. In: CINFORM, 5., 2004, Salvador. **Anais eletrônicos...** Salvador: UFBA, 2004. Disponível em: <http://www.cinform.ufba.br/v_anais/artigos/asafujino.html>. Acesso em: 10 ago. 2006.

GARCÍA GUTIERREZ, A.; LUCAS FERNÁNDEZ, R. Lenguajes documentales e información de actualidad. In: ____. **Documentación automatizada en los medios informativos**. Madrid: Paraninfo, 1987. cap. 3, p.67-90.

GENIEVA, E. Access to information and "public domain" in the post - 'Perestroyka'. Russia: a paradoxal experience. In: INFOETHICS, 3., 2000, Paris. **Anais eletrônicos**… Paris: UNESCO, 2000. Disponível em: <http://webworld.unesco.org/infoethics2000/documents/paper_genieva.rtf>. Acesso em: 1 abr. 2006.

GIACAGLIA, M.E. A organização da informação em sistemas CAD: análise crítica de esquemas existentes e proposta para o caso brasileiro. **Sinopses,** São Paulo, v.35, p.70-74, 2001.

GOMES, H. E. **Informação Científica**. Disponível em: <http://academica.extralibris.info/biblioteconomia/informacao_cientifica_hagar_es.html>. Acesso em: 16 ago. 2006.

GONZALEZ DE GÓMEZ, M.N.G. de; CANONGIA, C. (Org.). **Contribuição para políticas de ICT**. Brasília: IBICT, 2001.

GRUPO STELLA. Plataforma Lattes. Disponível em: <http:// www.stela.ufsc.br/legado/revistaplataformalattes.pdf >. Acesso em: 18 maio 2007.

GUARINO, N. Understanding, building, and using ontologies: a commentary to "Using Explicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga. **International Journal of Human and Computer Studies**, v.46, p. 293-310, 1997.

GUEDES, V.L.S.; BORSCHIVER, S. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. In: CINFORM, 6., 2006, Salvador. **Anais eletrônicos...** Salvador: UFBA, 2006. p.1-18. Disponível em: <http://www.cinform.ufba.br/vi_anais/docs/VaniaLSGuedes.pdf>. Acesso em: 03 ago. 2007.

HUTCHINS, W.J. **Languages of indexing and classification**: a linguistic study of structures and functions. Herts: Peter Peregrinus, 1978. (Librarianship and Information Studies, 3).

IBICT. **Guia de fontes de financiamento à Ciência & Tecnologia.** 6. ed. Brasília: CNPq/IBICT, 1993. 197 p.

INGERWERSEN, P. **Information Retrieval Interaction**. London: Taylor Graham, 1992. Disponível em: <http://www.db.dk/pi/iri>. Acesso em 03 abr. 2003.

IYER, H. **Classificatory Structures:** Concepts, Relations, and Representation. Frankfurt: Verlag, 1995. 229 p.

KENT, A. **Manual da recuperação mecânica da informação.** São Paulo: Polígono, 1972. 427 p.

KOBASHI, N.Y.; SANTOS, R.N.M. dos. Institucionalização cognitiva da pesquisa científica no Brasil sob a ótica da Ciência da Informação. **Journal of the American Society of Information Science**, 2007. [No prelo].

KOBASHI, N.Y.; SMIT, J.W.; TÁLAMO, M. de F.G.M. A função da terminologia na construção do objeto da Ciência da Informação. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v.2, n.2, abr. 2001. Disponível em: <http://www.dgzero.org/abr01/art_03.htm>. Acesso em: 31 jul. 2002.

KURAMOTO, H. Biblioteca Digital Brasileira: integrando a ICT brasileira. In: MARCONDES, Carlos Henrique et al. **Bibliotecas Digitais:** saberes e práticas. 2. ed. Salvador: EDUFBA; Brasília:IBICT, 2006. Cap. 5, p. 287-303.

LANCASTER, F.W.(1979). **Information Retrieval Systems**: characteristics, testing and evaluation. 2.ed. Nova York: John Wiley & Sons, 1979.

_____ .(2004). **Indexação e resumos**. 2.ed. rev. atual. Brasília: Briquet de Lemos/Livros, 2004.

LE COADIC, Y. **A Ciência da Informação**. 2.ed. Brasília: Briquet de Lemos/Livros, 2004.

LIMA, V. M. A. **Da classificação do conhecimento científico aos sistemas de recuperação de informação:** enunciação de codificação e enunciação de decodificação da informação documentária. 2004. 155 f. Tese (Doutorado em Ciências da Comunicação) - Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 2004.

LIMA-MARQUES, M. **Ontologias**: da filosofia à representação do conhecimento. Brasília: Thesaurus, 2006. 72 p. (Ciência da Informação e da Comunicação, 1).

LOPES, I. L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ciência da Informação,** Brasília, v.31, n.1, p.41-52, jan./abr. 2002.

MACHADO, A.M.N. **Informação e controle bibliográfico**: um olhar sobre a cibernética. São Paulo: Editora UNESP, 2003.

MANDER, R.; SALOMON, G.; WONG, Y. A "pile" metaphor for supporting casual organization of information. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 24., 1992, California. **Proceedings...** California: ACM Special Interest Group On Computer-human Interaction, 1992. p. 627 - 634.

MARCONDES, C. H. Metadados: descrição e recuperação de informações na web. In: MARCONDES, C. H. et al. **Bibliotecas Digitais:** saberes e práticas. 2. ed. Salvador: EDUFBA;Brasília: IBICT, 2006. Cap. 2, p. 95-111.

MARCONDES, C. H.; MENDONÇA, M. A. R.; MALHEIROS, L. R. A estrutura dos elementos de metodologia científica no texto de artigos científicos em ciências da saúde. In: CONGRESSO MUNDIAL DE INFORMAÇÃO EM SAÚDE E BIBLIOTECAS, 9., 2005, Salvador. **Anais eletrônicos...** Salvador: ICML, 2005. Disponível em: <http://www.icml9.org/program/track5/public/documents/Carlos%20H-181056.pdf>. Acesso em: 13 ago. 2006.

MARCONDES, C. H.; SAYÃO, L. F. (2001). Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, Brasília, v. 30, n. 3, p.24-33, set./dez. 2001.

_____. (2001). Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. **Ciência da Informação**, Brasília, v. 3, n. 31, p.42-53, set./dez. 2002.

MARQUES, P. **Modernização do Brasil**: dilemas e perspectivas. 2.ed. São Paulo : IEA/USP, 1994. 53 p. (Ciência e Tecnologia, 18).

MARTINS, E. V. O contexto político e o discurso da ciência da informação no Brasil: uma análise a partir do Ibict. **Ciência da Informação**, Brasília, v. 33, n. 1, p.91-100, jan./abr. 2004.

MATHIAS, P. Who unbound Prometheus? In: Mathias, Peter (ed.). **Science and Society 1600-1900**. Cambridge: Cambridge University Press, 1972, p.54-79.

MEADOWS, A. J. (1990). Theory in Information Science. **Journal of Information Science**, v. 16, p.59-63, 1990.

_____. (1999). **A comunicação científica**. Brasília: Briquet de Lemos/Livros, 1999.

MÉNDEZ RODRÍGUEZ, E. **Metadados y recuperación de información:** estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: Trea, 2002.

MOREIRA, A. **Tesauros e Ontologias**: estudo de definições presentes na literatura das áreas das Ciências da Computação e da Informação, utilizando-se o Método Analítico-Sintético. 2003. 150 f. Dissertação (Mestrado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2003.

MOREIRA, A.; ALVARENGA, L.; OLIVEIRA, A. de P. O nível do conhecimento e os instrumentos de representação: tesauros e ontologia. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v. 5, n. 6, dez. 2004. Disponível em: <http://www.dgz.org.br/dez04/Art_01.htm>. Acesso em: 27 mar. 2006.

NORUZI, A.(2006). Folksonomies: (Un)Controlled Vocabulary? **Knowledge Organization**, v.33, n.4, p.199-203., 2006.

_____ . (2007). Folksonomies: Why do we need controlled vocabulary? **Webology**, v.4, n.2, Jun. 2007. Disponível em: < http://www.webology.ir/2007/v4n2/editorial12.html >. Acesso em: 17 ago. 2007.

PACHECO, R. C. dos S.; KERN, V. M. Uma ontologia comum para a integração de bases de informações e conhecimento sobre ciência e tecnologia. **Ciência da Informação**, Brasília, v. 30, n. 3, p.56-63, set./dez. 2001.

PACKER, A. et al. SciELO: uma metodologia para publicação eletrônica. **Ciência da Informação**, Brasília, v.27, n.2, p.109-121, maio/ago. 1998.

PINHEIRO, L.V.R. Comunidades científicas e infra-estrutura tecnológica no Brasil para uso de recursos eletrônicos de comunicação e informação na pesquisa. **Ciência da Informação**, Brasília, v. 32, n. 3, p.62-73, set./dez. 2003.

PINHEIRO, L.V.R; LOUREIRO, J.M.M. Traçados e limites da Ciência da Informação. **Ciência da Informação**, Brasília, v.24, n.1, p. 42-53, jan./abril 1995.

PINTO, G.R.P.R.; PEREIRA, H.B. De B.; BURNHAM,T.F. Definição de uma ontologia para os canais preferenciais de difusão do conhecimento técnico-científico: fase de preparação. In: CINFORM, 6., 2005, Salvador. **Anais...** Salvador: UFBA, 2005. 1 CD-ROM.

POMBO, O. **Da classificação dos seres à classificação dos saberes**. Disponível em: <http://www.educ.fc.ul.pt/hyper/resources/opombo-classificacao.pdf>. Acesso em: 29 set. 2007.

RADA, R. Focus on links: a holistic view of hypertext. **International Classification**, v.18, n.1, p. 13-18, 1991.

RAYWARD, W.B. Some schemes for restructuring and mobilising information in documents: a historical perspective. **Information Processing & Management**, v.30, n.2, p.163-175, 1994.

RIOS, J. A. Ontologias: alternativa para a representação do conhecimento explícito organizacional. In: CINFORM, 6., 2005, Salvador. **Anais...** Salvador: UFBA, 2005. 1 CD-ROM.

RÍOS, R. de los; SANTANA, P. H. de A. El espacio virtual de intercambio de información sobre recursos humanos en Ciencia y Tecnología de América Latina y el Caribe Del CV Lattes al CvLAC. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 42-47, set./dez. 2001.

ROBREDO, J. **Documentação de hoje e de amanhã**: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. rev. e ampl. Brasília: Edição do autor, 2005.

ROWLEY, J. **A biblioteca eletrônica**. 2. ed. Brasília: Briquet de Lemos/Livros, 2002.

SANTANA, P.H. de A. et al. Servidor de enlaces: motivação e metodologia. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 48-55, set./dez. 2001.

SANTOS, P.R.E. dos. **Arquivos de cientistas**: gênese documental e procedimentos de organização. São Paulo: Associação dos Arquivistas de São Paulo, 2005.

SAYÃO, L..F. Bases de dados: a metáfora da memória científica. **Ciência da Informação**, Brasília, v.25, n.3, p.232-240, 1996.

SENA, N. K. Open archives: caminho alternativo para a comunicação científica. **Ciência da Informação**, Brasília, v.29, n.3, p.71-78, set./dez. 2000.

SILVA, F. M. e. Análise da Revista Ciência da Informação disponibilizada na Scielo a partir do seu vocabulário controlado. **Transinformação**, Campinas, v. 14, n. 2, p.133-138, jul./dez. 2002.

SILVA, G. L. da. A política da União Européia no domínio da informação científico-tecnológica. **Ciência da Informação**, Brasília, v. 26, n. 1, p.72-77, jan./abr. 1997.

SILVA, M. R. da. **Análise bibliométrica da produção científica docente do programa de pós-graduação em Educação Especial/UFSCar: 1998-2003**. 2004. 177 f. Dissertação (Mestrado em Educação Especial) - Universidade federal de São Carlos, São Carlos, 2004.

SMIT, J. W.; KOBASHI, N. Y**. Como elaborar vocabulário controlado para a aplicação em arquivos**. São Paulo: Arquivo do Estado, Imprensa Oficial do Estado de São Paulo, 2003. (Como fazer, 10).

SMIT, J. W.; KOBASHI, N. Y.; TÁLAMO, M. de F. G. M. A determinação do campo científico da Ciência da Informação: uma abordagem terminológica. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v. 5, n. 1, fev. 2004. Disponível em: <http://www.datagramazero.org.br/fev04/Art_03.htm>. Acesso em: 20 maio 2006.

SONDERGAARD, T.F.; ANDERSEN, J.; HJØRLAND, B. Documents and the communication of scientific and scholarly information: Revising and updating the UNISIST model. **Journal of Documentation**, v.59, n.3, p.278-320, 2003.

SOUZA, M.I.F.; VENDRUSCULO, L.G.; MELO, G.C. Metadados para a descrição de recursos de Informação em meio eletrônico: utilização do padrão Dublin Core. **Ciência da Informação**, Brasília, v.29, n.1, p.93-102, jan./abril 2000.

SOUZA, R. R. Sistemas de Recuperação de Informações e Mecanismos de Busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.11, n.2, p.161-173, maio/ago.2006.

SOUZA, R. R ; ALVARENGA, L.. A web semântica e suas contribuições para a Ciência da Informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abril 2004.

SOWA, J.F. **Conceptual Structures**: Information processing in mind and machine. Massachusetts: Addison-Wesley Publishing, 1984. (System Programming Series).

SVENONIUS, E. **Intelectual foundation of Information Organization.** Cambridge: Mit Press, 2001.

TÁLAMO, M. de F.G.M. **Linguagem documentária**. São Paulo: APB, 1997. 12p. (Ensaios APB, 45).

TARGINO, M. das G. Novas Tecnologias e Produção Científica: uma relação de causa e efeito ou uma relação de muitos efeitos?. **DataGramaZero – Revista de Ciência da Informação**, Rio de Janeiro, v. 3, n. 6, dez. 2002. Disponível em: <http://www.dgzero.org/dez02/Art_01.htm>. Acesso em: 03 ago. 2006.

TARGINO, M. das G.; GARCIA, J. C. R. Ciência brasileira na Base de Dados do Institute for Scientific Information (ISI). **Ciência da Informação**, Brasília, v.29, n.1, p.103-117, jan./abr. 2000,

TOUTAIN, L. M. B. B. Biblioteca digital: definição de termos. In: MARCONDES, C. H. et al. **Bibliotecas Digitais:** saberes e práticas. 2. ed. Salvador: EDUFBA; Brasília: IBICT, 2006. p.18-19.

TRISKA, R.; CAFÉ, L. Arquivos abertos: subprojeto da Biblioteca Digital Brasileira. **Ciência da Informação**, Brasília, v. 30, n. 3, p.92-96, set./dez. 2001.

TRISTÃO, A. M. D. et al. Sistema de classificação facetada: instrumento para organização da informação sobre cerâmica para revestimento. **Informação & Sociedade: Estudos**, João Pessoa, v. 14, n. 2, p.1-18, 2004. Disponível em: <http://www.informacaoesociedade.ufpb.br/ojs2/index.php/ies/article/view/62/60>. Acesso em: 11 maio 2006.

VALENTIM, M. L. P. Informação em ciência e tecnologia: políticas, programas e ações governamentais – uma revisão de literatura. **Ciência da Informação**, Brasília, v. 31, n. 3, p.92-102, set./dez. 2002.

VAN RIJSBERGEN, C. J. **Information Retrieval.** 2. ed. London: Butterworths, 1979.

VAN SLYPE, G. **Los lenguajes de indización**: concepción, construcción y utilización en los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez, 1991. (Biblioteca del Libro).

WEISMAN, H.M. **Information Systems, services and centers**. New York: Becker & Hayes, 1972.

WERSIG, G. Information Science: the study of postmodern knowledge usage. **Information Processing and Management**, v.29, n.2, p.229-239, 1993.

ZIMAN, J. **Conhecimento público**. Belo Horizonte: Itatiaia; São Paulo: EDUSP, 1979. (O Homem e a Ciência, 8).

ANEXO

A PLATAFORMA LATTES[68]

A Plataforma Lattes representa a experiência do CNPq no que se refere à integração de seus sistemas de informações gerenciais, instrumento fundamental não só para as atividades de fomento operadas pela Agência, mas também para tratamento e difusão das informações necessárias à formulação e à gestão de políticas de ciência e tecnologia.

O CNPq vem buscando integrar as suas bases de informações. Essa integração tem como fonte primária de coleta de dados de quatro projetos distintos, porém integrados, são eles: Sistema Eletrônico de Currículos, o Diretório dos Grupos de Pesquisa no Brasil; o Diretório de Instituições; e o Sistema Gerencial de Fomento.

O primeiro deles (e o que interessa a nossa pesquisa) se refere a um Sistema Eletrônico de Currículos. O registro da vida pregressa e atual dos pesquisadores é elemento fundamental para a análise de seu mérito e competência. Nesse domínio, o Brasil logrou desenvolver um formato-padrão para coleta de informações curriculares, adotado não só pelo CNPq, mas pela maioria das agências de fomento do País.

Fazem uso desse sistema pesquisadores, estudantes, gestores, profissionais e demais atores do sistema nacional de Ciência, Tecnologia e Inovação.

No CNPq, suas informações são aplicadas: na avaliação da competência de candidatos à obtenção de bolsas e auxílios; na seleção de consultores, de membros de comitês e de grupos assessores; no subsídio à avaliação da pesquisa e da pós-graduação brasileiras.

Histórico do Currículo Lattes

De 1993 a 1999, o CNPq utilizou formulários em papel, sistema em ambiente DOS (BCURR) e sistema de currículos específicos para credenciamento de orientadores (MiniCurrículo). Nesse período, a Agência acumulou cerca de 35 mil registros curriculares da atividade de C&T do País. Embora os instrumentos tenham viabilizado a operação de fomento da Agência, a natureza das informações dificultava a completa utilização dessa operação em outros processos de gestão em C&T (por exemplo, não era possível separar co-autores ou mesmo contabilizar índices de co-autoria nos currículos).

Entre 1998 e 1999, o CNPq realizou levantamento junto à comunidade de consultores ad hoc visando estabelecer um modelo de currículo que atendesse tanto às suas necessidades de operação de fomento como de planejamento e gestão em C&T. Além disso, o grupo de desenvolvimento "Grupo Stela" incluiu no formulário eletrônico diversas funcionalidades há muito solicitadas pela comunidade científica, tais como relatórios configuráveis, saída para outras fontes, indicadores de produção, dicionários individualizados, importação dos dados preenchidos em outros sistemas de currículos, etc.

Entre março e abril de 1999, 140 dos 400 consultores que responderam à pesquisa avaliaram o primeiro protótipo do currículo Lattes (à época denominado CV-Genos). Em maio de 1999, CNPq e Capes acordaram

[68] Extraído de http://lattes.cnpq.br/conheca/con_apres.htm

completa compatibilização do novo currículo do CNPq com os dados de pós-graduação, sob a ótica dos indivíduos de um Programa (pesquisadores, docentes ou discentes). O encontro entre as agências resultou na modificação do protótipo, que se transformou no Sistema de Currículos Lattes e foi lançado a 16 de agosto de 1999.

Nos dois primeiros anos do Sistema de Currículos Lattes, a cobertura de currículos ligados a C&T aumentou em mais de 300%, com a base anterior de cerca de 35.000 registros sendo incrementada para mais de 100 mil currículos.
Interação com outras bases de C&T

Em julho de 2000, a Coordenação Geral de Informática do CNPq iniciou um trabalho de intercâmbio com outras instituições ligadas a C&T no País. O resultado foi a ligação dinâmica dos currículos Lattes do CNPq com referência ao mesmo pesquisador em outras bases de dados. Ao mesmo tempo que construiu o formulário off-line, a Coordenação Geral de Informática do CNPq também trabalhou na ferramenta on-line, que funciona sobre uma plataforma Web e permite que os pesquisadores atualizem os seus currículos diretamente na base do CNPq.

Nesse trabalho de intercâmbio, o CNPq vinculou os currículos Lattes com: INPI, para apresentação dinâmica das patentes de registro dos pesquisadores; com SCIELO, LILAC, MEDLINE (acordo com a BIREME), para leitura dos textos completos publicados pelos pesquisadores (e para vínculo com os currículos dos co-autores); com as universidades, para vínculo com bases institucionais desses pesquisadores.

No ano de 2000, as Instituições Federais de Ensino Superior reuniram suas equipes de informática no Workshop de Sistemas de Informações das IFES (UFOP - Ouro Preto) e convidaram as agências federais para construção de um modelo único de informação, visando racionalizar o processo de captura de dados no Sistema Federal de Educação em Ciência e Tecnologia.

Na ocasião, o CNPq prontificou-se a construir projeto específico para atender a essa demanda, mas salientou a necessidade de manter a confidenciabilidade das informações (e a Plataforma operacional) dos pesquisadores, dado que estas são o principal subsídio ao processo de fomento.

Em fevereiro de 2001, UFSC, UNICAMP, UFRJ, USP, UFRGS, UFBA e UFRN, universidades que haviam procurado o CNPq solicitando abertura tecnológica de sua plataforma, participaram de workshop na Agência, visando à construção da Linguagem de Marcação da Plataforma Lattes (LMPL), sob coordenação da CGINF/CNPQ, sendo os trabalhos de desenvolvimento conduzidos pelo Grupo Stela da UFSC.

Desse encontro, resultou a formação da Comunidade Virtual LMPL, que definiu o modelo DTD (Data Type Definition) XML do Currículo Lattes, que faz parte da versão 1.4. Com esse modelo, as universidades brasileiras podem extrair informações do currículo Lattes e/ou gerar informações para o mesmo a partir dos seus sistemas corporativos. O projeto viabilizou a abertura da Plataforma Lattes, do ponto de vista de conteúdo dos dados, e manteve inalterado o acesso técnico às informações, preservando a segurança dos pesquisadores.

Em julho de 2000, a BIREME promoveu um encontro em São Paulo, no qual o CNPq foi convidado a mostrar sua experiência com a Plataforma Lattes. Nesse encontro, estavam representantes dos Conicyts do Chile, da Venezuela

e do México, e da Organização Pan-Americana de Saúde. O CNPq apresentou o Diretório dos Grupos de Pesquisa e o site de acesso ao Sistema de Currículo Lattes, o que despertou o interesse da Organização Pan-Americana de Saúde, que construiu um formulário latino-americano, denominado CvLAC, a partir da experiência do currículo brasileiro. O Grupo Stela foi contratado para esse fim, dando início aos trabalhos em fevereiro de 2001, e o CNPq disponibilizou a Plataforma gratuitamente para que o projeto alcançasse âmbito latino-americano.

Em abril de 2001, aconteceu uma grande conferência, estando presentes mais de 500 pessoas, entre as quais representantes dos Conicyts e representantes de bibliotecas virtuais, principalmente do Scielo. O CNPq apresentou todo o histórico de construção da Plataforma Lattes. A partir daí, o projeto chamou a atenção não só das áreas de saúde dos países latino-americanos que já tinham o reconhecimento a partir da Organização Pan-Americana de Saúde mas também da própria operação do Conicyt.

Números da Plataforma Lattes até outubro de 2005:
- 604.395 currículos enviados ao CNPq;
- 19.470 grupos de pesquisa cadastrados;
- 335 instituições cadastradas;
- 77.649 pesquisadores cadastrados;
- 47.973 pesquisadores doutores cadastrados.

Intercâmbio de dados

A Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior (CONSCIENTIAS) foi criada para desenvolver ontologias que se prestem ao intercâmbio de informações entre agências de fomento e instituições ligadas ao tema Ciência, Tecnologia, Inovação e Informações de Aprimoramento de Nível Superior. Uma ontologia é usada para indicar um domínio de conhecimento ou o domínio semântico para uma unidade de informação.

Na CONSCIENTIAS, as ontologias são representadas pela linguagem de marcação XML (e**X**tensible **M**arkup **L**anguage) e têm por finalidade principal o estabelecimento de uma forma comum de troca de informações entre agências de fomento e suas instituições usuárias.

Caracterizam-se como responsabilidades da Comunidade CONSCIENTIAS a concepção, elaboração, recomendação e manutenção das gramáticas relacionadas às ontologias submetidas pelas agências ou instituições conselheiras.

A referida Comunidade é uma extensão da Comunidade LMPL (Linguagem de Marcação da Plataforma Lattes), estabelecida no ano 2000 para ser responsável pela criação e manutenção das gramáticas XML da Plataforma Lattes. Sua criação coroa o processo de aproximação entre agências federais e estaduais, em um movimento de padronização de informações e racionalização de procedimentos, envolvendo fornecimento e intercâmbio de informações em benefício das comunidades científicas, tecnologias e de educação superior.

A definição do padrão de currículos Lattes em XML significa para as instituições de ensino e pesquisa um intercâmbio de informações curriculares entre as suas bases institucionais e as bases do Sistema de Currículo Lattes.

Para essas instituições, a adoção do padrão nacional definido pela Comunidade CONSCIENTIAS-LMPL de exportação e importação de currículos garante segurança e estabilidade nas regras de tradução entre suas estruturas de dados e a estrutura do Currículo Lattes.

a) Ontologias recomendadas

São definidas como ontologias recomendadas aquelas que já foram submetidas à análise, avaliadas, criticadas e testadas pelos grupos técnicos das instituições conselheiras da Comunidade CONSCIENTIAS-LMPL.

A ontologia passa por essas etapas em que os conselheiros, através do portal da Comunidade, submetem suas críticas e sugestões ao padrão que está sendo recomendado. Em conjunto os conselheiros determinam os prazos para avaliação e submissão das críticas até chegarem a um acordo.

As ontologias apresentadas aqui já foram discutidas e estão aptas à adoção por qualquer instituição que queira trocar informações entre os instrumentos da Plataforma Lattes e seus sistemas corporativos.

b) Padronização XML: Curriculum Vitae

O padrão XML para o Curriculum Vitae foi a primeira unidade de informação definida para a Plataforma Lattes. Esse padrão mantido pela Comunidade CONSCIENTIAS-LMPL foi elaborado seguindo as informações e a estrutura delas representadas no Sistema de Currículos Lattes.

Através da definição feita pela Comunidade CONSCIENTIAS-LMPL, para a unidade de informação de Currículo Vitae, o sistema de Currículos Lattes incorporou as funcionalidades de integração de suas informações em XML, sendo disponibilizadas a partir da versão 1.4 deste sistema.

Esse padrão XML foi inicialmente construído utilizando a linguagem de definição de tipos, DTD (Document Type Definition). Posteriormente, com a homologação da linguagem XML Schema pelo Consórcio W3C, a comunidade CONSCIENTIAS-LMPL construiu uma nova gramática utilizando a linguagem de esquemas para o mesmo padrão XML de Currículo Vitae.

Com essa linguagem, o XML Schema, pode-se utilizar de recursos anteriormente não disponíveis na linguagem antecessora - o DTD, como mecanismo de controle de tipos, a utilização de namespaces, e a reutilização de código.