Qualidade de bases de dados para construção de indicadores de C&T: a produção científica do CETEM e o Currículo Lattes

Jackson de Figueiredo Neto

Universidade Federal do Rio de Janeiro
Escola de Comunicação

Mestrado em Ciência da Informação
Convênio UFRJ/ECO – MCT/IBICT

Orientador:
Profa. Maria de Nazaré Freitas Pereira
Doutora em Ciências Humanas, IUPERJ

Rio de Janeiro
2003

Qualidade de bases de dados para construção de indicadores de C&T: a produção científica do CETEM e o Currículo Lattes

Jackson de Figueiredo Neto

Dissertação submetida ao curso de Mestrado da Pós-Graduação em Ciência da Informação do MCT/IBICT em convênio com a UFRJ/ECO, como parte dos requisitos necessários ao grau de Mestre.

Aprovada por:

_____
Profa. Maria de Nazaré Freitas Pereira – Orientador
Doutora em Ciências Humanas, IUPERJ. Rio de Janeiro.

_____
Profa. Rosali Fernandez de Souza
PhD, Polytechnic of North London / CNAA, Inglaterra.

_____
Profa. Maria Luiza Machado Campos
Doutora em Information Systems, University Of East Anglia, UEA, Grã-Bretanha.

_____
Profa. Hagar Espanha Gomes
Livre Docência. Universidade Federal Fluminense, UFF, Niteroi.

_____
Prof. Geraldo Moreira Prado
Doutor em Desenvolvimento, Agricultura e Sociedade, UFRRJ, Itaguaí.

Rio de Janeiro
2003

Figueiredo Neto, Jackson de

Qualidade de bases de dados para construção de indicadores de C&T: a produção científica do CETEM e o Currículo Lattes / Jackson de Figueiredo Neto. – Rio de Janeiro : UFRJ/ECO, 2003.

177p.

Dissertação (mestrado em Ciência da Informação). Universidade Federal do Rio de Janeiro/ECO - MCT/IBICT, 2003.

1. Bases de dados 2. Qualidade 3. Indicadores de C&T.
I. Título

Dedicatória

À Denise e à Julia,
duas estrelas que iluminam meu caminho
com a luz da alegria.

Agradecimentos

À minha orientadora, Profa. Maria de Nazaré Freitas Pereira, pela competência e objetividade que tanto contribuíram para a concretização deste trabalho.

Ao Departamento de Ensino e Pesquisa do IBICT. Aos professores, pela minha formação, aos funcionários, pelo apoio. Em especial, aos professores e professoras Gilda Olinto, Rosali Fernandez de Souza, Lena Vania R. Pinheiro e Alexandre Pedrini pela dedicação e entusiasmo demonstrados ao longo das aulas das disciplinas cursadas.

À equipe de bibliotecárias do CETEM, Ana Oliveira, Sonia Mamede e Jacira Coutinho, pela participação dedicada e profissional na localização e recuperação dos documentos utilizados no trabalho de campo da presente dissertação.

À equipe de colegas do Serviço de Informação do CETEM - SEIN, Alexandre Prado, Andréa Vilhena, Carlos Campos, Carlos David, Carlos Souza, César Silva, Maurício Pinheiro, Paulo Sérgio Costa e Vera Ribeiro, pelo profissionalismo com que se dedicam às suas atividades, tendo me proporcionado a tranqüilidade necessária para levar a cabo essa empreitada.

À Diretoria do CETEM, em especial, aos meus superiores durante o período do curso de mestrado, Dr. Fernando Lins, Dr. Juliano Barbosa, Dr. Gildo Sá (*in memoriam*) e Dr. Augusto Wagner, que me propiciaram todo apoio institucional para a realização do mesmo.

Aos amigos e familiares que me acompanharam e me apoiaram durante o mestrado, em especial, à Maria Helena L. Hatschbach, à Ana Carolina P. D. da Fonseca e ao Camilo Figueiredo Fernandes.

Aos meus queridos pais, Maria Gilda e Luiz Alves de Figueiredo (*in memoriam*), pelo exemplo de vida.

Obrigado.

Resumo

FIGUEIREDO NETO, Jackson de. **Qualidade de bases de dados para construção de indicadores de C&T:** a produção científica do CETEM e o Currículo Lattes. Orientador: Maria de Nazaré Freitas Pereira. Rio de Janeiro: Universidade Federal do Rio de Janeiro, Escola de Comunicação – MCT/IBICT, 2003. 177p. Dissertação. (Mestrado em Ciência da Informação).

O presente trabalho analisa os principais aspectos que envolvem a qualidade de bases de dados para a produção de indicadores de C&T. É apresentado um breve histórico do desenvolvimento dos indicadores de C&T e de sua utilização. Faz-se uma revisão bibliográfica dos conceitos, métodos e sistemas da qualidade aplicados às bases de dados. Apresenta-se uma metodologia com o objetivo de avaliar a qualidade dos dados da base Currículo Lattes como fonte primária para a construção de indicadores de C&T precisos e confiáveis.

Abstract

FIGUEIREDO NETO, Jackson de. **Qualidade de bases de dados para construção de indicadores de C&T:** a produção científica do CETEM e o Currículo Lattes. Orientador: Maria de Nazaré Freitas Pereira. Rio de Janeiro: Universidade Federal do Rio de Janeiro, Escola de Comunicação – MCT/IBICT, 2003. 177p. Dissertação. (Mestrado em Ciência da Informação).

The present work analysis the main aspects related to the quality of the database for the production of S&T indicators. A brief background on the development of S&T indicators and their use is presented as well as a bibliographic review of the concepts, methods and quality systems applied to databases. A methodology is proposed to evaluate the quality of the "Curriculo Lattes" database as a primary source for the establishment of precise and dependable S&T indicators.

Lista de Siglas

ABIPTI	Associação Brasileira dos Institutos de Pesquisa Industrial
ABNT	Associação Brasileira de Normas Técnicas
ARIST	Annual review of information science and technology
BIRD	International Bank for Reconstruction and Development
BLR & DD	British Library Research and Development Department
Capes	Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CD-ROM	disco ótico de dados
C&T	Ciência e Tecnologia
CenDoTeC	Centro Franco-Brasileiro de Documentação Técnica e Científica
CETEM	Centro de Tecnologia Mineral
CIQM	Centre for Information Quality Management
CL	Currículo Lattes
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CONSER	Cooperative Online Serials
CPRM	Companhia de Pesquisa de Recursos Minerais
CVLAC	Currículo Vitae em Ciência e Tecnologia
DESIRE	Development of a European Service for Information on Research and Education
EUSIDISC	European Association of Information Services
FMI	Fundo Monetário Internacional
HTML	Hiper Text Marked Language
HTTP	Hiper Text Transport Protocol
IBBD	Instituto Brasileiro de Bibliografia e Documentação
IBGE	Instituto Brasileiro de Geografia e Estatística
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
IETF	Internet Engineering Task Force
IMO	Information Market Observatory
ISBN	International Standard Book Number Number
ISI	Institute for Scientific Information
ISSN	International Standard Serial

KDD	Knowledge Discovery in Databases
LC	Library of Congress
LILACS	Literatura Latino-Americana e do Caribe em Ciências da Saúde
MARC	Machine-readable bibliographic records
MCT	Ministério da Ciência e Tecnologia
MEC	Ministério da Educação e Cultura
NSF	National Science Foundation
OCDE	Organização para Cooperação e Desenvolvimento Econômico
OCLC	Online Computer Library Center
OMS	Organização Mundial da Saúde
ONU	Organização das Nações Unidas
OPAS	Organização Pan-Americana de Saúde
OST	Observatoire des Sciences et des Techniques
PCC	Program for Cooperative Cataloging
RLIN	Research Library Network
SCI	Science Citation Index
SciELO	Scientific Eletronic Library Online
SCOUG	Southern California User Group
SEBRAE	Serviço Brasileiro de Apoio às Micros e Pequenas Empresas
SPROU	Science Policy Research Unit
TI	Tecnologias de Informação
TQM	Total Quality Management
UNESCO	United Nations Educational, Scientific and Cultural Organization
UKOLN	UK Office for Library and Information Networking
XML	Extensible Markup Language
WLN	Washington Library Network
WorldCat	OCLC Online Union Catalog
WWW	World Wide Web

Lista de Quadros

Lista de Anexos

Sumário

INTRODUÇÃO

Nos tempos atuais, em que o ambiente informacional, além de complexo e sofisticado, disponibiliza uma quantidade fenomenal de informação, os indivíduos e as organizações tendem, cada vez mais, a depositar mais confiança nos sistemas de informação apoiados em bases de dados eletrônicas do que nas suas experiências diretas. Por outro lado, constata-se que, normalmente, o usuário não-especialista de uma base de dados não costuma questionar sobre os aspectos que definem o conteúdo de uma base de dados, como, por exemplo, se a mesma foi concebida para atender a um tipo de demanda específica ou como o produtor daquela base de dados controla a precisão e a atualização dos dados que são nela alimentados. Tal atitude pode comprometer seriamente a qualidade das decisões tomadas a partir de informações extraídas dessas bases de dados, trazendo como efeito imediato o descrédito do próprio sistema.

As causas deste não questionamento podem ser as mais variadas, desde a simples ignorância sobre estes aspectos, passando por um certo conformismo em não mexer nesse "enxame de abelhas" que tal iniciativa representa, até a falta de alternativas de outras fontes de informação. De uma maneira ou de outra, todo usuário de base de dados, sabe, em maior ou menor grau, do abismo que existe entre aceitar candidamente o que uma base de dados oferece e questionar objetivamente o conteúdo da mesma. Em outras palavras, verifica-se que existe uma incapacidade do usuário não-especialista em saber avaliar objetivamente a qualidade desse produto informacional denominado base de dados.

Mais especificamente, o presente trabalho traz essa importante discussão para a área de gestão de C&T na qual a utilização de indicadores, construídos a partir de bases de dados, vem ganhando importância significativa ao longo da última década. A comunidade científica vem exigindo, cada vez mais, processos de tomada de decisão mais transparentes, baseados em regras claras e menos subjetivismos provenientes de decisões tomadas por meia dúzia de "iluminados" em gabinetes fechados. Tais exigências convergem para a necessidade de sistemas de informação mais robustos e confiáveis. Portanto, esse conhecimento quantitativo, representado pelos indicadores, adquire uma relevância crescente na medida em que os governos e as instituições

caminham no sentido de, não apenas atender essas exigências mais imediatas, mas, também, atender às condicionantes econômicas do mundo atual - qualidade, competitividade e produtividade.

Esta pesquisa é parte integrante do Projeto CNPq - *Por uma Economia do Conhecimento: Avaliação de Bases de Dados Nacionais para a Produção de Indicadores de C&T (Ciência e Tecnologia)*, coordenado pela orientadora desta dissertação. A pesquisa explora a componente qualidade de bases de dados cadastrais para a produção de indicadores de C&T, contribuindo para a realização do último módulo do projeto e, conseqüentemente, para sua conclusão[1].

Conforme informa Pereira[2], as bases bibliográficas nacionais podem ser de dois tipos: as bases de produção científica originadas a partir do controle das publicações científicas de grupos de pesquisa, apresentando forte orientação institucional e as bases bibliográficas originadas do controle da literatura científica, principalmente a periódica, apresentando forte orientação temática. No Brasil, os levantamentos realizados demonstram que é crescente o surgimento de bases de controle da produção científica e que as tradicionais bibliografias brasileiras tendem a se extinguir.

Portanto, com o declínio das bases bibliográficas brasileiras surgem questões acerca da substituição destas pelas bases cadastrais. Por apresentarem métodos de produção distintos, as bases cadastrais parecem oferecer limitações em comparação às bases bibliográficas tradicionais, principalmente quando se deseja a partir delas (das bases cadastrais) extrair determinados tipos de informação que pudessem atender às atuais demandas dos usuários destas bases como, por exemplo, para a construção de mapas de conhecimento e estudos estratégicos de C&T.

No Brasil, verifica-se na literatura diversas iniciativas no sentido de se criar um sistema nacional de informações em C&T. Uma das primeiras iniciativas foi a criação do IBBD

---

[1] PEREIRA, Maria de Nazaré Freitas. **Por uma Economia do Conhecimento**: Avaliação de Bases de Dados Nacionais para a Produção de Indicadores de C&T (Ciência e Tecnologia). Relatório Parcial (Avaliação de qualidade de bases de dados bibliográficas). Rio de Janeiro, julho/2001. Processo 520416/93-7 (NV).

[2] PEREIRA, M. N. F. *et al.* Bases de dados na economia do conhecimento: a questão da qualidade. **Ciência da Informação**, Brasília, v.28, n. 2, 1999. p. 1. Disponível em: http://www.ibict.br/cionline/280299/28029913.htm. Acesso em: nov. 2002.

em 1954, diretamente subordinado ao CNPq. Era o órgão no Brasil que apoiava a pesquisa e promovia o acesso à informação técnico-científica no país e no exterior. Em 1975 foi criado o SNDCT – Sistema Nacional de Desenvolvimento Científico e Tecnológico com o objetivo de tornar disponíveis informações sobre C&T. Para tal, foram necessárias mudanças institucionais para se permitir a formulação de uma política de informação. Tais mudanças culminaram com a extinção do IBBD em 1976 e a criação do IBICT. Este novo órgão teria funções mais amplas do que o antigo IBBD, apoiando as ações do SNDCT, sob a coordenação do CNPq. A implantação do SNICT – Sistema Nacional de Informação em C&T estabelecia como objetivo a formação de uma rede nacional de cooperação e intercâmbio para assegurar o aproveitamento integral dos conhecimentos adquiridos no país e no exterior. O SNICT não chegou a ser implantado. Resultado disso, criou-se um vácuo nas funções que um dia pertenceram ao extinto IBBD.

Na década de 80, através do PBDCT, tornou-se possível a elaboração de documento de ação programada, elaborado pelo CNPq, em informação em C&T. Duas iniciativas merecem destaque: o subprograma do PADCT em informação e tecnologia, sob a responsabilidade do IBICT e o Plano de Biblioteca Universitárias, elaborado pelo Ministério da Educação.

Apesar dos esforços que se concretizam mais ao nível do planejamento do que de sua implementação, o Brasil dispõe de incipiente infra-estrutura de informação bibliográfica, materializada em bases de dados, excetuando-se áreas de Medicina, Agricultura e Nuclear.[3]

No início da década de 90, registra-se um projeto desenvolvido pelo CNPq que tinha como objetivo a construção de um sistema de informação sobre a atividade científica e tecnológica no âmbito de universidades e institutos de pesquisa, com cobertura nacional. Segundo Guimarães[4], as raízes deste projeto surgem a partir de uma demanda do então Secretário de C&T (1990), José Goldemberg, que encomenda um

---

[3] BATTAGLIA, M. G. B. **Análise sistêmico documental e proposta de um sistema de informação em C&T para a FINEP**. Rio de Janeiro: UFRJ, Escola de Comunicação – CNPq/IBICT, 1992. 112p. Dissertação. (Mestrado em Ciência da Informação).

[4] GUIMARÃES, R. **Avaliação e fomento de C&T no Brasil: propostas para os anos 90**. Brasília: MCT/CNPq, 1994. 178p. p. 112.

levantamento de grupos de pesquisa em atividade no país de forma permitir a criação de "mapas" para a orientação na montagem de um programa de apoio aos "laboratórios associados". Portanto, o projeto do CNPq assimila esta abordagem e define o "grupo de pesquisa" como sua unidade de análise, apresentando como principal justificativa o fato da unidade de análise apresentar a possibilidade de apreensão do modo pelo qual se organiza o processo de produção do conhecimento. Além disso, o "grupo de pesquisa" como unidade de análise permitiria uma adequação à crescente interdisciplinaridade observada na pesquisa científica.

Esse projeto cuja etapa de implementação iniciou-se em 1992, denominou-se *Diretório de Grupos de Pesquisa no Brasil*. Seu principal objetivo é de constituir e manter atualizada uma base de dados censitária sobre a atividade de pesquisa no país, através do registro da composição e das atividades dos grupos de pesquisa ativos.[5] Posteriormente, o Diretório integrou-se a outros sistemas de informação. Este conjunto de sistemas passou a denominar-se Plataforma Lattes.

Concebida para integrar os sistemas de informações das agências federais de financiamento das atividades de C&T, racionalizando o processo de gestão de C&T, a Plataforma Lattes foi lançada em agosto de 1999, com a disponibilização do sistema Currículo Lattes à comunidade de pesquisadores do país. Ao longo dos últimos anos, novos sistemas e aperfeiçoamentos têm sido incorporados à Plataforma Lattes, visando sua consolidação como principal subsídio à tomada de decisão em Ciência, Tecnologia e Inovação do sistema de C&T nacional e sua integração às ações de intercâmbio de informações e de subsídios à formação de comunidades virtuais temáticas, em âmbito internacional.[6]

A Plataforma Lattes tem como objetivo a compatibilização e a integração das informações coletadas em diferentes momentos de interação do CNPq com seus usuários. Atualmente a Plataforma Lattes engloba quatro sistemas que operam de forma integrada:

- Sistema de Currículos Eletrônicos (Currículo Lattes),

[5] GUIMARÃES, R. *op.cit.*
[6] CNPq. Plataforma Lattes. Disponível em: http://lattes.cnpq.br/. Acesso em: fev. 2003.

- Diretório dos Grupos de Pesquisa,
- Diretório de Instituições e
- Sistema Geral de Fomento.

Outras bases que não pertencem ao CNPq, como o SciELO, a LILACS, a base de patentes do INPI e os bancos de teses e dissertações das Universidades também fazem parte da Plataforma Lattes.

A partir do ano de 2000, o sucesso da Plataforma Lattes extrapolou as fronteiras do país, chegando ao conhecimento de autoridades internacionais em políticas de C&T e desencadeando uma série de acordos entre o CNPq e organismos de C&T de outros países. O primeiro foi o convênio estabelecido entre o CNPq e a Organização Pan-Americana de Saúde - OPAS, que propiciou a tradução do sistema Currículo Lattes para o espanhol, na forma do Sistema CVLAC (*Currículo Vitae em Ciência e Tecnologia*). Dessa forma, a OPAS estará disponibilizando a metodologia do Currículo Lattes para os organismos de C&T dos países da América Latina e do Caribe. Posteriormente, o CNPq foi convidado a desenvolver acordos bilaterais de cooperação com a Colômbia para a constituição do programa piloto "Diretório Latino-Americano em Ciência & Tecnologia", que permitiu a consulta conjunta, por palavras em português e espanhol, à pesquisa brasileira e colombiana. Este programa piloto despertou o interesse de outros países como o Chile, México e Portugal. Portanto, a importância e a abrangência da Plataforma Lattes não se limita mais ao país, vai além, no sentido de se tornar o portal da produção científica dos países de língua espanhola e portuguesa. Neste aspecto, a Plataforma Lattes poderá vir a preencher a lacuna existente na base do ISI cuja cobertura é deficiente em relação à produção bibliográfica dos países em desenvolvimento.

Considerando que os sistemas da Plataforma Lattes trabalham de forma integrada e de que os dados primários mais importantes para a derivação de indicadores de produção são os dados de cada pesquisador individual, indicando sua origem, sua formação e sua produção científica, a presente Dissertação definiu como objeto de estudo um desses sistemas que compõem a Plataforma Lattes, o Sistema de Currículos Eletrônicos ou, como é mais conhecido, sistema Currículo Lattes.

Portanto, o objetivo do presente trabalho é avaliar o grau de precisão e confiabilidade dos dados contidos no Currículo Lattes e, por conseguinte, sua adequação como fonte primária de dados para a construção de indicadores de C&T precisos e confiáveis.

A presente Dissertação foi escrita seguindo uma estrutura de conteúdo que permitisse ao leitor, inicialmente, familiarizar-se com o contexto do uso dos indicadores de C&T e das bases de dados, em especial, as bases bibliográficas. Em seguida, faz-se uma revisão bibliográfica dos conceitos, métodos e sistemas da qualidade aplicados às bases de dados. Finalmente, apoiado nessa base teórica, é proposta uma metodologia com o objetivo de avaliar a qualidade da base Currículo Lattes como fonte primária para a construção de indicadores de C&T.

Dessa forma, o capítulo 1, "Indicadores de C&T", descreve o significado e a importância dos indicadores no contexto das atividades de gestão em C&T, sua utilização como instrumento imprescindível para a formulação de políticas e para a avaliação e o acompanhamento das atividades de C&T de países, regiões e instituições.

O capítulo 2, "Bases de dados", apresenta os principais conceitos que envolvem o produto base de dados e situa sua importância no contexto da indústria da informação. É dado destaque às bases bibliográficas que representam a produção científica.

O capítulo 3, "Qualidade em bases de dados", aborda as bases de dados a partir da ótica da qualidade. Apresenta uma revisão dos conceitos e métodos aplicados ao tema qualidade. Aborda as técnicas para avaliação da qualidade de bases de dados, critérios de controle de qualidade e apresenta exemplos de métodos de controle de qualidade aplicados em bases bibliográficas.

O capítulo 4, "Material e Método", apresenta a base Currículo Lattes como objeto de estudo e descreve a metodologia proposta para a avaliação da mesma. São descritos os critérios utilizados para a definição das amostras de dados retiradas da base Currículo Lattes e submetidas à avaliação.

O capítulo 5, "Resultados", apresenta os resultados obtidos com a aplicação da metodologia.

Finalmente, o capítulo 6, “Considerações Finais”, discute os resultados obtidos, aponta possíveis causas para os problemas encontrados e apresenta algumas soluções no sentido de melhorar o nível de qualidade dos dados contidos na base Currículo Lattes.

1. INDICADORES DE C&T

1.1. Histórico e importância

Facilitado pelo desenvolvimento das tecnologias de informação, o uso de indicadores tornou-se um fenômeno típico das últimas quatro décadas. As atividades humanas ficaram mais fáceis de serem monitoradas e estudadas e o uso de indicadores tem permitido melhor planejamento das políticas públicas. Essa quantidade fenomenal de informação que hoje o mundo dispõe, quando bem utilizada, propicia melhorias nos processos de tomada de decisão de governos e empresas e na qualidade de vida da população.

Para citar alguns exemplos, além dos indicadores econômicos tradicionais como o PIB – Produto Interno Bruto, PNB – Produto Nacional Bruto, Renda *per capita* e inflação, os estudiosos da área contam hoje com o IDH – Índice de Desenvolvimento Humano, IPH – Índice de Pobreza Humana, Índice de Evasão Escolar, dentre outros. Importantes organismos internacionais dedicam-se à tarefa de acompanhamento desses indicadores. Destacam-se entre eles: Organização das Nações Unidas – ONU, *International Bank for Reconstruction and Development* – BIRD, *United Nations Educational, Scientific and Cultural Organization* – UNESCO, Organização para Cooperação e Desenvolvimento Econômico – OCDE, *National Science Foundation* – NSF, Fundo Monetário Internacional – FMI e Organização Mundial da Saúde – OMS.

No Brasil, a Fundação Instituto Brasileiro de Geografia e Estatística – IBGE é a instituição responsável pela elaboração e levantamento dos principais indicadores sociais e econômicos do país. Além do IBGE, alguns Ministérios e outras instituições, como a Fundação Getúlio Vargas, acompanham alguns indicadores isolados.[7]

Quanto aos indicadores de C&T, objeto de maior interesse desta Dissertação, eles adquirem uma importância vital nos processos de decisão dos governos pela percepção generalizada de que a pesquisa científica e tecnológica tornou-se atividade essencial para a geração de riquezas e a promoção do bem estar social.

---

[7] PINTO, M. M. N. **Indicadores de P&D do setor produtivo no Brasil:** situação, necessidades e perspectivas. Orientador: Paulo César Gonçalves Egler. Brasília: Universidade de Brasília, Centro de Desenvolvimento Sustentável, 2000. 74p. Dissertação. (Mestrado em Desenvolvimento Sustentável).

O desenvolvimento científico e tecnológico é uma das metas fundamentais da política científica, principalmente no que diz respeito aos índices e padrões de desenvolvimento econômico, seu direcionamento e seus efeitos sociais. É através da formulação de políticas que as nações direcionam suas atividades de C&T como meta para atingir os objetivos nacionais. Assim, nenhuma nação atinge os objetivos de desenvolvimento em todos os aspectos sem uma efetiva infra-estrutura em C&T, consolidada através de uma política de informação.

Segundo Barré,

> *indicadores de C&T são conhecimento quantitativo sobre os parâmetros da atividade científica, tecnológica e de inovação a níveis institucional, disciplinar, setorial, regional, nacional e plurinacional. Tal conhecimento tem como objetivo caracterizar e posicionar instituições, regiões ou países em 'mapas' temáticos, permitindo, assim, o estudo comparativo, incluindo análise sobre o tempo*.[8]

Portanto, os indicadores de C&T permitem acompanhar em níveis e aspectos diversos a dinâmica das atividades científicas e tecnológicas e efetuar comparações. A correta compreensão desta dinâmica, destes movimentos e das forças presentes, exige dados quantitativos robustos[9], de preferência reunidos em bases de dados eletrônicas que facilitam sobremaneira o estabelecimento de determinadas relações entre seus registros. As relações destes registros e entre registros de bases distintas funcionam como a matéria-prima que permite a construção de indicadores.[10]

As primeiras experiências no campo de estudos dos indicadores de C&T se dão nos Estados Unidos, Inglaterra e França na década de 70. Posteriormente, nas décadas de 80 e 90 outros países da Europa e América Latina também desenvolveram estudos nesta área. Tais estudos abordam aspectos relacionados com a concepção teórica de indicadores, suas metodologias de trabalho, o formato organizacional e os produtos. Entre os estudos pioneiros destaca-se a primeira edição do relatório "*Science and Engineering Indicators*", publicado nos Estados Unidos pela *National Science Foundation* (NSF) em 1972. Na Inglaterra merece menção o "*Science Policy Research*

---

[8] Barré *apud* PEREIRA, M. N. F. *et al*. Bases de dados na economia do conhecimento: a questão da qualidade. **Ciência da Informação**, Brasília, v.28, n. 2, 1999. Disponível em: http://www.ibict.br/cionline/280299/28029913.htm. Acesso em: nov. 2002.

[9] MUSTAR, P. **Les chifres clés de la science & de la technologie**. Ed. 1998-1999, Paris: OST, 1998. 111p. p.5.

[10] PEREIRA, M. N. F. *op. cit*.

Unit" (SPROU). Na França, os primeiros estudos foram realizados através de um organismo internacional, a Organização para a Cooperação e o Desenvolvimento Econômico (OCDE) que preparou uma série de estudos quantitativos para subsidiar especialistas na formulação de políticas de C&T dos países signatários. Atualmente, o conjunto de indicadores usados pelos países da OCDE constitui-se na principal fonte de referência para o desenvolvimento de indicadores relacionados às atividades científicas e tecnológicas.

No Brasil, na área de Ciência e Tecnologia, os indicadores mais importantes são atualmente levantados pelo Ministério de Ciência e Tecnologia - MCT. Com a sua criação em 1985, o MCT passou a assumir a responsabilidade pela organização e divulgação das informações de C&T do país, de forma centralizada. Para realizar essa tarefa, o MCT conta com a colaboração de inúmeras instituições dos governos federal e estaduais, assim como organizações privadas que produzem informações para construção de indicadores de C&T.[11]

O Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq foi a instituição que realizou os primeiros esforços para construir indicadores de C&T para o país. Nos anos 80, o CNPq iniciou a coleta e a publicação de informações sobre os recursos do Governo Federal aplicados em C&T, fazendo uso de recomendações do Manual Frascati da OCDE e orientações da UNESCO. Outras instituições envolvidas na construção e no desenvolvimento de estudos sobre o tema e que merecem menção, são: o Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT e a Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes).[12]

Para o Brasil, assim como para todos os países em desenvolvimento, os desafios relacionados aos indicadores de C&T não se limitam apenas às questões de escopo e cobertura destes, mas também, a melhorar a qualidade das informações primárias através do desenvolvimento de estudos visando conhecer as estruturas de C&T e aperfeiçoar as metodologias utilizadas na produção de indicadores.

---

[11] APRESENTAÇÃO / histórico. **Indicadores de C&T**. Disponível em: http://www.mct.gov.br/estat/ascavpp/portugues/menu1page.htm. Acesso em: jul. 2003.

[12] APRESENTAÇÃO / histórico. *op. cit*.

Que tipos de decisões devem ser tomadas com base nos indicadores de C&T? Por que precisamos desenvolver indicadores estratégicos? Eles têm algum significado? O que eles nos dizem? Com essas questões, Kondo chama a atenção sobre as reais finalidades do uso dos indicadores de C&T. Ele observa que, infelizmente, alguns produtores e usuários destes indicadores tendem a considerar esses números *como representantes de algum tipo de "verdade" sobre o estado da ciência e da tecnologia, e não como possíveis aproximações da realidade*.[13] O que o autor quer enfatizar é que os indicadores têm suas limitações e que, além do mais, *os indicadores somente serão úteis se forem confiáveis*. Vale lembrar que essas considerações estão em sintonia com o objetivo central desta Dissertação, qual seja, dar uma contribuição para a área de estudos de indicadores no sentido de explorar e ilustrar questões sobre qualidade dos dados primários utilizados para a construção de indicadores de C&T.

Portanto, para responder à questão sobre quais seriam as razões para desenvolver indicadores de C&T, Kondo propõe que, ao se construir indicadores confiáveis, deve-se considerar as seguintes razões para usá-los:

1. *Permitir uma melhor compreensão sobre a contribuição do progresso técnico ao crescimento econômico.*

2. *Ajudar a responder a perguntas sobre políticas de C&T.*

3. *Realizar as seguintes funções: monitorar o desempenho do sistema de C&T, avaliar e modificar a alocação de recursos para melhorar a eficiência do sistema de C&T, justificar ou negociar os orçamentos de C&T e oferecer insumos para o estabelecimento de políticas de C&T.*

4. *Apoiar as seguintes atividades: prestação de assessoria a ministros e a outros altos funcionários, prestação de contas aos contribuintes, análise do sistema nacional de inovações.*[14]

---

[13] KONDO, E. K. Desenvolvendo indicadores estratégicos em ciência e tecnologia: as principais questões. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 128-133, maio/ago 1998. p. 128.
[14] KONDO, E. K. *op. cit*. p. 129.

1.2. Indicadores de insumos e produtos

Para melhor compreender este tópico, é importante, antes de abordar as características dos indicadores de insumo e de produtos, esclarecer três conceitos distintos: atividade, produtividade e progresso.[15]

Atividade. São os esforços e a energia despendidos em determinada tarefa sem levar em consideração se eles foram aplicados de maneira adequada ou não.

Produtividade. Significa o grau em que essas atividades produzem resultados relevantes.

Progresso. Mede o grau em que a produtividade nos leva aos resultados desejados.

Portanto, para aqueles que se utilizam das análises quantitativas da ciência, a medição dos insumos indica a atividade, e a medição dos produtos indica a produtividade. Por outro lado, encontrar formas de medir o progresso é uma tarefa bem mais difícil, ocorrendo muitas divergências entre aqueles que se dedicam a esse tema.

1.2.1. Indicadores de insumos

No início do desenvolvimento dos indicadores de C&T, os levantamentos se limitavam ao dimensionamento dos recursos financeiros e humanos investidos em C&T. Não por acaso os indicadores de insumos são os que possuem séries históricas mais longas e mais detalhadas. Nos países em desenvolvimento, a maioria dos estudos se concentra nos indicadores de insumo. As razões são óbvias. Em primeiro lugar, o levantamento desses indicadores não é uma tarefa sofisticada e, ao se examinar planos de desenvolvimento, verifica-se que tais planos servem para vender a imagem de um país dedicado à pesquisa científica e os indicadores de insumos mostram essa atividade.

O número de pessoas dedicadas à atividade científica, o número de instituições, a capacitação dos pesquisadores, recursos financeiros destinados a uma determinada área da ciência, são alguns exemplos de indicadores de insumos. Esses indicadores, embora

---

[15] VELHO, L. Indicadores científicos: aspectos teóricos y metodológicos. In: MARTINEZ, E. (ed.). **Ciencia, tecnologia y desarrollo: interrelaciones teóricas y metodológicas**, Caracas: Nueva Sociedad, 1994. p. 307-348. p. 310.

apresentem limitações diversas, são importantes para a elaboração de políticas setoriais e mesmo para a avaliação e acompanhamento dos indicadores de produtos.[16] Problemas metodológicos prejudicam comparações internacionais, pois nem todos os países aplicam a mesma metodologia para o levantamento desses indicadores. Raras vezes a qualidade é considerada nas medições de insumos. Por isso, deve-se considerar os indicadores de insumos como indicadores da atividade científica e não de produtividade ou de progresso.[17]

Os indicadores de insumos podem ser desagregados segundo três dimensões[18]:

- a natureza da pesquisa: básica, aplicada e atividades correlatas.

- os setores que executam ou financiam estas atividades: governo, universidades, empresas (simplificadamente).

- a classificação dos recursos de cada um destes setores, obedecendo critérios específicos para o governo (segundo objetivos sócio-econômicos), as instituições de nível superior (segundo áreas do conhecimento) e as empresas (segundo setores de atividade econômica).

1.2.2. Indicadores de produtos

A intangibilidade dos produtos da ciência torna mais difícil a quantificação da atividade científica, ou ainda melhor, da produtividade científica. Assim, os produtos da ciência (conhecimentos e idéias) em vista da dificuldade em medi-los, exige que tais medições devam se realizar de forma indireta, principalmente através dos indicadores bibliométricos, considerando duas premissas básicas:

1. A meta central da ciência é a produção de novos conhecimentos.

2. O produto da ciência se reflete totalmente nos instrumentos de escrita formais dos cientistas, em especial, nos periódicos científicos.

---

[16] PINTO, M. M. N. *op. cit*. p. 25.
[17] VELHO, L. *op. cit*. p. 311.
[18] APRESENTAÇÃO / histórico. *op. cit*.

Tais premissas estão fundamentadas na tradição mertoniana. Segundo Merton, o objetivo da ciência é *a ampliação do conhecimento científico certificado* e o pesquisador que se dedica a outras metas, como, por exemplo, buscar a solução de problemas práticos, este é visto como "periférico" à profissão.[19]

Porém, outra vertente do pensamento afirma que a publicação formal é apenas um tipo de comunicação científica e, mais ainda, ela seria menos significativa do que os meios informais. Segundo Velho, existem outras metas da atividade científica: a contribuição para a solução de problemas práticos, a transmissão de uma perspectiva científica a toda a população de um país, a educação de especialistas, o aumento do prestígio nacional e internacional, etc.[20] Portanto, partindo dessa linha de pensamento, as medidas quantitativas seriam apenas indicadores parciais da atividade científica.

Além dos problemas conceituais dos indicadores de produto, existem as dificuldades metodológicas, principalmente aquelas relacionadas à construção e ao tamanho das bases de dados. Nos capítulos seguintes dessa Dissertação estas dificuldades serão abordadas em detalhe.

Apesar de todas essas questões, ainda é através dos indicadores quantitativos que se medem os produtos científicos, sendo os mais importantes: o número de autores científicos, o número de publicações e a contagem de citações. Inicialmente os indicadores de produtos, também chamados indicadores de resultados, limitavam-se à produção científica. Posteriormente, foram incorporadas a produção de patentes e a transferência de tecnologia entre países.

Além da sua utilização para comparações internacionais e na formulação das políticas de C&T, os indicadores de produtos são utilizados para identificar:[21]

1. a evolução das atividades científicas e tecnológicas nos países, nas indústrias e nas sociedades.

2. a mudança de estrutura tecnológica e o avanço tecnológico.

---

[19] Merton *apud* VELHO, L. *op. cit*. p.312.
[20] VELHO, L. *op. cit*. p.312.
[21] PINTO, M. M. N. *op. cit.* p.25.

3. a dependência, a difusão e penetração da tecnologia.

4. a produtividade dos trabalhos científicos e tecnológicos e os impactos das novas tecnologias.

1.2.2.1. Indicadores da produção científica

Dentre os inúmeros indicadores de produtos de C&T, é de particular interesse para a presente Dissertação um conjunto de indicadores que mede os resultados da atividade científica, os chamados indicadores bibliométricos. Estes serão alvo do estudo de caso apresentado nas páginas seguintes do presente trabalho.

Os indicadores bibliométricos são utilizados desde o início do século passado. Tais indicadores são obtidos a partir de análises estatísticas dos dados quantitativos obtidos da literatura técnico-científica. Os trabalhos de Bradford, Zipf, Lotka e outros estudiosos permitiram demonstrar que a literatura científica tem a propriedade de mostrar um comportamento estatístico regular[22]. Em 1965, Price relacionou o crescimento do conhecimento científico com o aumento dos documentos publicados e formulou a lei do crescimento exponencial da ciência.[23] A partir daí, iniciou-se a aplicação de métodos científicos para analisar a própria ciência. Pritchard definiu o termo "bibliometria" como o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. Segundo Macias-Chapula, *a bibliometria desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar a tomada de decisão*.[24]

Price demonstrou que todas as distribuições bibliométricas se ajustam a distribuições hiperbólicas de "vantagem cumulativa", o que significa dizer que, por exemplo, quanto mais trabalhos um determinado autor produz, mais facilidade ele terá em produzir

---

[22] SANCHO, R. Indicadores bibliométricos utilizados em la evaluacion de la ciência y la tecnologia, revision bibliográfica. **Revista Española de Documentación Cientifica**, Madrid, v. 13, n. 3-4, p. 842 –865, 1990. p. 845.
[23] SANCHO, R. *op.cit*. p. 844.
[24] MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 134-140, maio/ago 1998. p.134.

outros, ou, quanto mais citações ele recebe, maior possibilidade ele terá de ser citado, isto é, o êxito gera mais êxito.[25]

Sancho[26] chama a atenção para dois fenômenos importantes ocorridos na década de 60 que muito contribuíram para o grande número de estudos bibliométricos realizados naquele período. O primeiro deles foi a informatização das bases de dados que facilitou enormemente a tarefa da pesquisa dos dados, e, o segundo, foi o aumento da demanda pelos órgãos governamentais de estudos de avaliação da eficácia de suas políticas em C&T.

Na década de 70 um importante fato estabeleceu um novo marco na área de estudos de gestão de C&T, a comercialização das bases de dados do *Institute for Scientific Information* – ISI. Dentre outros produtos de informação comercializados pela empresa de Eugene Garfield, destaca-se a criação do *Science Citation Index* – SCI. Foi a primeira base de dados (e única até os dias atuais) de citação completa e sistemática em meio eletrônico, provocando uma revolução na maneira de avaliar a produtividade científica, passando a ser utilizada por diversas instituições como uma ferramenta inestimável para a política científica. A partir daí, foi possível sustentar a idéia de que a análise quantitativa da ciência passaria a ser uma ferramenta útil e confiável para a tomada de decisão de política científica.[27]

Não é intenção deste trabalho a descrição detalhada e a avaliação de todos os indicadores da produção científica. Entretanto, vale a pena destacar os mais conhecidos e mais utilizados nos estudos atuais. Segue abaixo uma breve descrição realizada por Macias-Chapula dos indicadores que o autor considera os mais importantes no cenário nacional e internacional:[28]

Número de trabalhos – Reflete os produtos da ciência, pela contagem dos trabalhos e pelo tipo de documento (artigos, livros, relatórios, etc.).

Número de citações – Reflete o impacto dos artigos ou assuntos citados.

---

[25] SANCHO, R. *op. cit*. p. 845.
[26] SANCHO, R. *op. cit*. p. 845.
[27] VELHO, L. *op. cit*. p. 319.
[28] MACIAS-CHAPULA, C. A. *op. cit*. p 135.

Co-autoria – Reflete o grau de colaboração na ciência em nível nacional e internacional.

Número de patentes – Reflete as tendências das mudanças técnicas ao longo do tempo e avalia os resultados dos recursos investidos em atividades de P&D. Esses indicadores determinam o grau aproximado da inovação tecnológica de um país.

Número de citações de patentes – Mede o impacto da tecnologia.

Mapas dos campos científicos e dos países – Auxiliam a localizar as posições relativas de diferentes países na cooperação científica global.

Vale destacar que, por muito tempo, a avaliação da ciência limitou-se aos indicadores de insumos. Atualmente, devido à crescente necessidade de justificar para a sociedade os investimentos destinados ao setor de C&T e a comprovada relação do avanço tecnológico com o desenvolvimento econômico e social, o foco das avaliações voltou-se para os indicadores da produtividade científica, isto é, a preocupação daqueles que definem a política cientifica está concentrada nos resultados dessa atividade. Sem esses indicadores seriam inconcebíveis as análises das políticas de C&T que hoje tais indicadores permitem realizar.

2. BASES DE DADOS

2.1. Conceitos, contexto e tipologia

Pouco mais de três décadas separam a indústria da informação, dominada completamente pela mídia impressa, das atuais redes eletrônicas mundiais. Os produtos de informação eletrônica, hoje, se espalham por quase todos os segmentos sociais economicamente ativos. Neste contexto, o surgimento e a popularização das bases de dados foram os fenômenos mais notáveis ocorridos nestas três décadas de idade da indústria da informação. Segundo Sayão, "*as atividades relacionadas ao ciclo de produção de bases de dados criaram as bases da indústria da informação eletrônica como hoje ela é conhecida*". [29]

Esse novo mercado, o mercado da informação eletrônica, alterou sobremaneira o funcionamento da sociedade a ponto de estabelecer um novo marco, dando início ao que se convencionou chamar de "era do conhecimento".

Pereira, ao abordar as definições para o termo "era do conhecimento", observa a existência de um duplo entendimento para esse conceito: de um lado, a definição tradicional que diz que esse tipo de economia funciona de forma intensiva com base em conhecimento oriundo da pesquisa científica; de outro lado, essa nova economia dá ênfase ao conhecimento sobre o conhecimento, devidamente organizado e explorado em bases de dados primárias ou de indicadores, fornecendo informações em tempo real para tomada de decisão, seja em investimentos governamentais, seja no monitoramento dos setores de produção da economia, apenas para citar alguns exemplos. Em síntese, nessa nova economia *as transações entre seus atores são, cada vez mais, mediadas por produtos e serviços de alto valor informacional, transportados por meios telemáticos, eletrônicos e computacionais*.[30] Nesse contexto, a base de dados eletrônica emerge como o produto mais importante e que melhor representa a indústria da informação nos dias atuais.

---

[29] SAYÃO, L. F. Bases de dados e suas qualidades. In: LUBISCO, N.; BRANDÃO, L. (Ed.). **Informação e Informática**. Salvador: EDUFBA, 2000.

[30] PEREIRA, M. N. F. *op. cit*.

O termo "base de dados" pode ser definido como um conjunto de informações organizado de acordo com alguma regra ou princípio. Um catálogo telefônico é uma base de dados. Ele está organizado alfabeticamente e pode estar na forma eletrônica ou não. O catálogo de uma biblioteca é também uma base de dados pois as informações estão organizadas segundo um sistema próprio de classificação. Organização é a palavra-chave de uma base de dados. Portanto, uma base de dados é qualquer coleção organizada de informações, embora, no uso atual do termo, esteja relacionada à informação na forma eletrônica.[31]

Uma base de dados é composta de registros. Normalmente um registro se refere a um item na base de dados. O registro é composto de campos que são elementos de informação individuais. O catálogo de uma biblioteca é um bom exemplo: o arquivo com suas fichas organizadas em ordem alfabética é a base de dados. Cada ficha armazenada nas gavetas é equivalente a um registro da base, isto é, cada ficha descreve um item bibliográfico através de campos pré-definidos como título, autor, assunto, data da publicação, etc.

Existem diferentes tipos de bases. A natureza do conteúdo é um fator determinante no desenvolvimento da interface de acesso de uma base de dados. Bases de dados de consulta são, geralmente, uma compilação de fatos e análises projetados para responder perguntas. Algumas bases apresentam um escopo bem definido. Outras são abrangentes como, por exemplo, a *Encyclopaedia Brittanica Online*. As bases podem conter informações bibliográficas na forma de breve descrição e/ou registros de texto completo de artigos, peças, vídeos, etc.[32]

Outra abordagem que permite entender o significado das bases de dados é através dos conceitos que estão por detrás da estrutura do fluxo de comunicação, entre a geração e a recepção do conhecimento e sua evolução até a comunicação eletrônica. As tecnologias da informação tornaram mais fácil o acesso ao conhecimento disponível sobre determinadas áreas e especialidades. Essa disseminação do conhecimento se concretiza via indexação que, por sua vez, lança mão de uma terminologia padronizada e

---

[31] GALE/ALISE bibliographic instruction support program. Farmington Hills: Gale, 2001. Disponível em: http://www.galegroup.com/pdf/customer_service/alise.pdf. Acesso em: dez. 2002. p. 19.

[32] GALE/ALISE bibliographic instruction support program. *op.cit*. p.12.

estruturada.[33] Nesse sentido, a terminologia adquire a função, dentre outras, de representação para transferir o conhecimento, isto é, a terminologia atua como um meio comunicativo. Garcia e Targino[34] afirmam que a terminologia é a peça-chave dos especialistas. Somente através da utilização dos termos é que se permite aos especialistas expressarem e comunicarem seus conhecimentos. A terminologia é a base do pensamento especializado e, a esse pensamento especializado, formando um conjunto organizado de informações ou de documentos, convencionou-se chamar de bases de dados.

Sayão, ao analisar as bases de dados no âmbito da produção científica mundial, traça um paralelo entre as formas de incorporação de conhecimento nas bases de dados e o conceito de memória coletiva.

> *Fazemos apelo aos testemunhos para fortalecer ou debilitar, mas também para completar o que sabemos de um evento do qual já estamos informados de alguma forma, embora muitas circunstâncias nos permaneçam obscuras.* [35]

Ao citar as palavras de Maurice Halbwachs contidas no livro "Memória Coletiva", Sayão chama a atenção para o fato de que aquelas palavras também exprimem o sentimento ou o estado de espírito do pesquisador no momento em que ele interroga uma base de dados à procura de informações que insiram seu trabalho de pesquisa na ciência feita pelo seu grupo. Sayão, ainda acrescenta: *"O seu próprio desejo (do pesquisador) de informação é absolutamente nebuloso, fazendo com que suas interrogações só consigam se realizar durante o ato de busca. O processo de interação com os conhecimentos armazenados na base de dados é que estabelece o foco da questão"*. Esse processo se insere completamente nas rígidas imposições do método científico, da natureza tribal e cumulativa da ciência na qual o pesquisador deve fundamentar sua questão sobre o que já foi estabelecido. Caso contrário, ele está condenado à rejeição e ao esquecimento e, o seu saber, ao descrédito.[36]

---

[33] TARGINO M. 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. p.103.
[34] TARGINO M. G.; GARCIA, J. C. R. *op.cit*. p. 104.
[35] SAYÃO, L. F. Bases de dados: a metáfora da memória científica. **Ciência da Informação**, Brasília, v. 25, n. 3, 1996. p. 314.
[36] SAYÃO, L. F. *op.cit*. p. 314.

O caráter cumulativo da ciência resulta em um corpo de conhecimento baseado no consenso. Esse corpo de conhecimento é representado pela literatura técnico-científica, fruto mais óbvio e mais facilmente sujeito à mensuração da atividade científica.[37] Apesar dos avanços alcançados com as atuais tecnologias de armazenamento em meio eletrônico, ainda não foi possível armazenar toda a literatura científica. Faz-se necessário, portanto, que esse conhecimento sofra um processo de tradução, de representação, transformando-se em metaconhecimento. Esse metaconhecimento ou conhecimento virtual é o conteúdo das bases de dados, que só existe em função da vinculação remota com algum conhecimento real.[38]

As considerações acima são muito apropriadas porque delimitam um tipo de base de dados de especial interesse para o presente trabalho, as bases bibliográficas. Elas representam a literatura técnico-científica e constituem a fonte primária para a construção de um conjunto de indicadores dos mais representativos em C&T.

Pereira destaca a importância das bases de dados, em especial as bases bibliográficas, no tocante à crescente utilização das mesmas na produção de indicadores de C&T e, mais recentemente, *para produzir estudos estratégicos de C&T, área de conhecimento que se organiza sob a denominação de inteligência competitiva.*[39]

Duas grandes linhas de trabalho, os estudos sociais de ciência e tecnologia e a gestão de C&T fazem uso intensivo tanto das bases bibliográficas como as não bibliográficas. Tais estudos se utilizam de indicadores construídos a partir das informações obtidas destas bases de dados. Portanto, as bases de dados bibliográficas ou, na sua ausência, as de natureza cadastral que incorporam referências bibliográficas, *permitem conhecer coletivamente o produto intelectual dos pesquisadores, bem como a tecedura da rede social em que se sustenta, por meio da construção de indicadores.*[40]

As facilidades de acesso proporcionadas pelo advento da Internet provocaram, a partir da segunda metade da década de 90, uma explosão na criação e no uso de bases de dados. A disponibilidade do acesso às bases de dados através das redes de comunicação,

---

[37] SAYÃO, L. F. *op.cit.* p. 315.
[38] SAYÃO, L. F. *op. cit*. p.315.
[39] PEREIRA, M. N. F. *op cit*.
[40] OECD 1996 *apud* PEREIRA, M. N. F. *op cit*.

em CD-ROM e em meios magnéticos passou a apresentar muitas vantagens em relação às fontes impressas. A principal delas foi a redução do tempo de disponibilização e atualização das informações, o que pode significar o acesso à informação desejada horas ou até dias antes de aparecer na forma impressa. Muitas bases de dados são atualizadas diariamente ou a cada minuto, o que faz com que muitas informações, atualmente, só estejam disponíveis na forma eletrônica. Outra característica oferecida pelas bases de dados eletrônicas é o maior poder de recuperação. Muitos provedores de informação permitem a realização de buscas simultâneas em até centenas de bases de dados ao mesmo tempo, com a possibilidade de uso de recursos de pesquisa sofisticados, como os operadores "booleanos", de proximidade e truncamento, para citar apenas alguns. Tais características conferem às bases de dados um extraordinário poder de facilidade, flexibilidade e rapidez na formulação de pesquisas e na obtenção de respostas. Outra importante vantagem quando comparada às fontes de informação impressa é a possibilidade de imprimir a informação desejada em formatos personalizados e pagar apenas pela informação de interesse em um dado momento, ao invés de se comprar uma base de referência impressa na sua totalidade, normalmente de custo elevado, podendo ser pouco utilizada e tornar-se desatualizada rapidamente. Portanto, verifica-se que, com a evolução das redes, o uso da informação eletrônica apresenta uma tendência gradual e crescente quanto a sua importância e volume na indústria da informação.[41]

As bases de dados destacam-se entre os principais produtos oferecidos na forma eletrônica pela indústria da informação. Inicialmente, as bases de dados eram armazenadas em computadores centrais e disponibilizadas para os usuários remotos através de redes de comunicação. Mais tarde, com o aumento da capacidade de armazenamento e a drástica redução de custos dos meios magnéticos e óticos, foi possível disponibilizar localmente as bases de dados. A partir da segunda metade da década de 90, conforme já mencionado anteriormente, ocorreu uma explosão no uso das bases de dados graças às facilidades de acesso proporcionadas pela expansão da Internet. Em razão desses fatores o número de bases de dados cresce continuamente. Em 1982 contabilizava-se cerca de 770 bases de dados. Na segunda metade dos anos 90 o

[41] CENDÓN, B. V. Bases de dados de informação para negócios. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 30-43, maio/ago 2002. p.31.

número de bases já alcançava cerca de 10 mil.[42] Segundo Choo[43], um terço das bases de dados existentes podem ser classificadas como bases de dados de informações para negócios.

Deve-se destacar que o sucesso das bases de dados deve-se às facilidades conquistadas a partir da disponibilização das informações no formato eletrônico, facilitando o trabalho do pesquisador que se utiliza de todas as vantagens propiciadas pela mídia eletrônica em relação às fontes impressas. A utilização das bases de dados eletrônicas permite ao pesquisador encontrar as informações de que necessita em poucos minutos. De outra forma, essa mesma pesquisa poderia levar dias caso recorresse a fontes impressas e dispersas em locais distintos. Além disso, a pesquisa em bases de dados permite encontrar determinadas informações que seriam quase que impossíveis de serem descobertas em fontes impressas, devido à limitação de seus pontos de acesso e a impossibilidade da busca por palavras no texto completo.[44]

As bases de dados são classificadas em três tipos principais: as bases de dados bibliográficas ou referenciais, as bases de dados de texto completo e as bases fatuais.

As bases de dados bibliográficas ou referenciais contêm registros bibliográficos que permitem ao usuário localizar uma publicação específica (um artigo de periódico, uma tese, um livro, um relatório de pesquisa, etc.). Além dos elementos informacionais que caracterizam uma referência bibliográfica, algumas bases podem também fornecer o resumo dos documentos.

As bases de texto completo contêm o documento completo. Com o avanço das tecnologias de armazenamento em meio eletrônico, a inclusão do texto completo passou a ser uma tendência das bases mais modernas. A vantagem óbvia desse tipo de base é o acesso imediato ao documento. Nos dias atuais é comum encontrar o documento no formato PDF, um tipo de formato eletrônico que reproduz fielmente o *layout* de uma página impressa contendo texto, gráficos e imagens.

---

[42] Williams *apud* CENDÓN, B. V. *op. cit*. p. 31.
[43] Choo *apud* CENDÓN, B. V. *op. cit*. p. 31.
[44] CENDÓN, B. V. *op. cit*. p.42.

As bases de dados fatuais fornecem respostas imediatas às questões formuladas. Tais questões não visam a obter como resposta uma bibliografia. Um grande número de bases de dados fatuais fornece informações numéricas tais como cotações de ações, índices de inflação, indicadores de C&T, etc.[45]

É importante destacar que, na prática, verifica-se uma tendência no sentido da ocorrência de bases híbridas, isto é, algumas bases de dados incorporam características dos vários tipos de bases já descritas. Um exemplo é a base cadastral Currículo Lattes que combina informação bibliográfica com os dados de experiência profissional dos pesquisadores.

Nos dias atuais, uma série de questões a respeito das bases de dados comerciais e a Internet começaram a surgir, especialmente nos últimos anos com a ocorrência da disseminação explosiva da Internet, disponível a dezenas de milhões de pessoas espalhadas pelos quatro cantos do planeta. Nesse contexto muitos viam a vasta quantidade de informação grátis na Internet como uma ameaça aos serviços comerciais de bases de dados. A respeito desse assunto, Cendón observa que esses dois segmentos, a Internet e as bases de dados comerciais, devem ser percebidos como fontes complementares de informação e acrescenta:

> *Cada uma dessas modalidades de fontes eletrônicas de informação tem seus pontos fortes. A Internet não tem paralelo no que diz respeito à quantidade e variedade de informações grátis e às publicações cinzentas, que envolvem não apenas a literatura efêmera que as bibliotecas tendem a não coletar, mas todo o segmento de publicações não oficiais ou quase-oficiais. Distingue-se ainda pela possibilidade da interatividade e pela facilidade de se estabelecerem contatos com fontes pessoais e organizacionais de informação. Por outro lado, a informação na Internet pode ser de acesso demorado, é desorganizada e caótica e pode ter sua autoridade contestada, enquanto a informação em bases de dados pode ser cara, mas é pontual, precisa, confiável e pode ser obtida com mais rapidez.* [46]

Em um primeiro momento, o surgimento da Internet como fonte alternativa de informação provocou um impacto nas empresas que comercializavam bases de dados. Entretanto, aos poucos, essas empresas souberam tirar proveito deste novo contexto.

---

[45] CENDÓN, B. V. *op. cit*. p. 34.
[46] CENDÓN, B. V. *op. cit*. p. 42.

Empresas que forneciam bases de dados em CD-ROM passaram a oferecer o acesso às suas bases através da Internet. De forma gradual os usuários via *Web* foram suplantando o número de usuários de bases em CD-ROM com a vantagem de que o acesso via *Web* eliminava o limite de espaço do CD-ROM, a desatualização das informações e permitia a integração de diversas bases.

2.2. Fontes de informação em C&T

As informações primárias utilizadas na construção dos indicadores de C&T são provenientes de uma variedade de fontes, cabendo, na maioria dos casos, aos órgãos governamentais a tarefa de organizar e sistematizar as informações sobre as atividades de produção e de disseminação de indicadores de C&T. O processo exige um razoável grau de interação com as diversas instituições, públicas e privadas, responsáveis pelas informações primárias, uma vez que estas informações são produzidas a partir de metodologias distintas para atender finalidades específicas dessas instituições.

Atualmente, milhares de produtores de bases de dados e de serviços de informação são representados por algumas dezenas de empresas que podem ser de dois tipos: as generalistas e as especializadas.[47] As empresas generalistas oferecem produtos diversificados, isto é, bases de dados de diferentes tipos e variedade de assuntos (p. ex., agricultura, engenharia, ciências sociais). As empresas especializadas focalizam um assunto específico, por exemplo, notícias e publicações da área jurídica .

Entre as empresas generalistas, uma das maiores e mais diversificadas é a *Dialog Corporation*. Esta empresa foi vendida em março de 2000 à *Thompson Corporation*, somando os serviços *Dialog*, *DataStar* e *Profound* aos que a *Thompson* já possuía, como o *Westlaw*, *Gale Group*, *Information Access Company* e o *Institute of Scientific Information* - ISI, tornando-se um gigante na produção e distribuição da informação.[48] O *Dialog* oferece mais de 600 bases com ênfase para o setor empresarial. A *DataStar* fornece acesso a mais de 350 bases com ênfase nas fontes européias.

[47] CENDÓN, B. V. *op. cit*. p. 32.
[48] CENDÓN, B. V. *op. cit*. p. 33.

Outras empresas generalistas, segundo Cendón, que merecem destaque, são: a OCLC que oferece mais de 70 bases de dados em artes e humanidades, negócios e economia, educação, engenharia, tecnologia e ciências em geral; a *H. W. Wilson Company* que produz várias bases bibliográficas em áreas de informação científica e de negócios, a *Silverplatter* que disponibiliza mais de 200 bases de informação para negócios e C&T, a *ProQuest* que fornece bases na área de notícias, administração, economia, teses e dissertações; e a *EBSCO Publishing* que oferece bases com texto completo de cerca de dois mil títulos de periódicos em negócios, C&T, inteligência empresarial, bancos, contabilidade e finanças.[49]

Entre as empresas especializadas, Cendón destaca as seguintes bases: *Factiva*, especializada em informações financeiras publicadas em revistas e jornais, a *Profound*, dedicada exclusivamente a fornecer acesso a bases de relatórios de pesquisas de mercado, análises econômicas de mais de 190 países, relatórios financeiros de mais de 4,5 milhões de empresas e notícias de 27 *newswires* globais, a *SkyMinder* fornece acesso a diversas bases de dados, agregando informações sobre dados financeiros de empresas, perfis de executivos, informações de crédito, indústrias e notícias e a *Lexis-Nexis* especializada em informações da área jurídica, fornecendo texto completo de um grande número de publicações corporativas e de revistas.[50]

Seguindo uma tendência atual do mercado globalizado, ocorre também, entre as empresas da indústria de informação, a formação de grandes conglomerados, mediante a fusão das mesmas. Um exemplo típico foi a fusão da *Thompson* e da *Dialog*, já comentado anteriormente.

Como veremos adiante, cada país adota uma metodologia própria para a tarefa de coleta, sistematização, homogeneização, construção e divulgação dos indicadores.

Nos Estados Unidos observa-se uma característica especial que é o domínio de instituições privadas no desenvolvimento de indicadores de C&T.[51] Esta característica deve-se, em grande parte, à existência do *Institute for Scientific Information* – ISI. Fundada na década de 50 por Eugene Garfield, a empresa, situada na Filadélfia, possui

[49] CENDÓN, B. V. *op. cit*. p. 32.
[50] CENDÓN, B. V. *op. cit*. p. 33.
[51] VELHO, L. *op. cit*. p. 318

uma base de dados reconhecida mundialmente como uma das mais importantes fontes de informação da publicação bibliográfica em âmbito internacional. A base de dados do ISI abrange, no total, 16 mil títulos de revistas, livros e anais de congressos internacionais nas áreas de ciências, ciências sociais, artes e humanidades. Desse total, deve-se destacar os mais de 8 mil títulos de periódicos científicos correntes indexados anualmente pela base ISI. Para cada artigo publicado nesses periódicos a base ISI registra os dados bibliográficos completos, incluindo resumos originais em inglês, os endereços dos autores e editores e as referências bibliográficas citadas em cada artigo.[52]

A base ISI processa anualmente cerca de 800 mil artigos científicos em mais de 100 campos científicos especializados. A partir da compilação destas informações oferece diversos produtos de informação, destacando-se os seguintes:

Current Contents - CC. Por mais de 40 anos tem fornecido aos pesquisadores dados bibliográficos e os índices de conteúdos dos principais periódicos científicos a nível mundial, atualizados diariamente na *Web*.

Who is Publishing in Science - WIPIS. Oferece uma lista com os nomes dos autores de artigos registrados no *Current Contents* em determinado ano.

Science Citation Index – SCI. Publicado desde 1961, contém informação proveniente das citações bibliográficas de todos os artigos de periódicos processados pelo ISI.

Conforme já anteriormente mencionado, o SCI merece destaque pelo avanço que proporcionou aos estudos de bibliometria. A criação do SCI transformou a literatura científica em uma fonte sistematizada e de fácil acesso para a análise quantitativa da ciência.

A comercialização da base SCI permitiu que outras empresas viessem a oferecer outros produtos de informação. Assim, a *Computer Horizon Incorporation* - CHI oferece novas abordagens de pesquisa combinando dados do SCI com os da MEDLINE. O *Center for Research Planning* – CRP compete com o próprio ISI no desenvolvimento de análises de co-citação.

---

[52] TESTA, J. A base de dados ISI e seu processo de seleção de revistas. **Ciência da Informação**, Brasília,v. 27, n. 2, p. 233-235, maio/ago 1998. p. 233.

Na França, com a criação do *Observatoire des Sciences et des Techniques - OST* em 1990*,* foi estabelecida uma nova concepção de construção de indicadores de C&T a partir de bases de dados de terceiros e na forma organizacional de reunir instituições interessadas na produção desses indicadores. Atualmente, o OST constitui-se de 14 instituições associadas. Destes associados, 7 são ministérios (da Pesquisa, da Defesa, da Indústria, da Economia, do Exterior, do Meio Ambiente e da Infra-estrutura), 5 são centros e institutos nacionais de pesquisa (CEA, CNS, CNRS, INSERM, INRA), a *France Télécom* e a Associação Nacional de Pesquisa Tecnológica (ANRT). Cada um deles está representado no Conselho Administrativo do OST e são responsáveis pela definição da orientação de seus trabalhos assim como fornecem os recursos humanos e financeiros para a consecução dos mesmos.[53]

O conceito de "observatório" nos leva a entender um dos aspectos originais do OST, qual seja: ele não coleta dados primários. Todo seu trabalho é desenvolvido a partir de bases de dados disponíveis no mercado.

Por outro lado, isto significa que se faz necessário um enorme esforço para contornar problemas de falta de normalização, de inexistência de campos de dados importantes, comparabilidade entre as informações de modo que se possa obter uma única base central relacional, a base do Observatório.

O principal produto do OST é uma edição bienal denominada "*Science & Thecnologie Indicateurs*". Trata-se de uma densa publicação de mais de 500 páginas contendo centenas de tabelas que apresentam dados comparativos das atividades de C&T na França sob os mais variados aspectos, assim como comparações internacionais, preponderantemente com os países da União Européia, Estados Unidos e Japão.

Na Espanha, a sistematização e divulgação dos indicadores de C&T é realizada pelo *Ministerio de Ciência y Tecnologia* através de sua *Secretaria de Estado de Política Científica y Tecnológica*. Anualmente é publicado um documento contendo os indicadores básicos de C&T com o objetivo de apresentar o esforço público e privado no desenvolvimento das atividades de C&T e a disponibilidade dos recursos. A

[53] OBSERVATOIRE DES SIENCES E DES TECHNIQUES. **Science & tecnologie: indicateurs 1998**. Paris: Econômica, 1998. 551p. p. 3.

publicação apresenta informação quantitativa de P&D e inovação, baseada em dados estatísticos provenientes de instituições oficiais, nacionais e internacionais, como o *Instituto Nacional de Estadística* (INE), a *Oficina Española de Patentes y Marcas*, CINDOC (CSIC), OCDE, EUROSTAT, etc.[54]

No Brasil existe um esforço direcionado para dotar o país com um sistema de informações sobre os recursos humanos e produtos de C&T mais abrangente e mais confiável. Um dos resultados mais expressivos deste esforço concretiza-se através do desenvolvimento da Plataforma Lattes, um conjunto de bases de dados mantido pelo CNPq o qual, pela sua importância atual, foi selecionado como objeto de estudo do presente trabalho.

[54] ESPANHA. Ministerio de Ciencia y Tecnologia. **Indicadores del sistema español de ciencia y tecnologia**. Madrid, 2000. 35 p. p. 3.

3. QUALIDADE EM BASES DE DADOS

3.1. Qualidade: definições, conceitos e modelos

A qualidade é tema considerado nos processos gerenciais desde os anos 30. Entretanto, os conceitos de qualidade tornaram-se amplamente aceitos somente após a Segunda Grande Guerra. Nesta ocasião, os gerentes norte-americanos aplicaram com sucesso estes novos conceitos na reestruturação das empresas japonesas, destacando-se como pioneiros nesta área Dewing, Juran e Ishikawa.[55]

Como não poderia deixar de ser, sua característica de aplicação tão ampla não permite que o conceito de qualidade apresente uma definição única e universal. Juran (1988) sugere que a qualidade deveria ser compreendida como "adequação ao uso" (*fitness for use*). Crosby (1979) define qualidade como "conformidade com os requisitos" (*conformance to requirements*). Essas definições colhidas na literatura mostram que a qualidade não pode ser definida simplesmente como um conceito abstrato de "excelência" mas que deve ser vista em relação às necessidades do usuário do produto final. Clark, Money e Tynan (1990) apresentam uma definição de qualidade como sendo *o quão consistentemente um produto ou serviço prestado atende ou excede as necessidades e expectativas dos consumidores*.[56]

A qualidade de um produto é normalmente definida sob aspectos distintos, no caso do mesmo ser um bem ou um serviço. Aspectos como confiabilidade, durabilidade, desempenho e estética são facilmente aplicáveis aos bens. Quanto à qualidade de um serviço, os modelos de avaliação são mais recentes e os critérios de qualidade mais difíceis de serem definidos devido à natureza intangível dos serviços.[57]

Quando um cliente/usuário adquire um produto ele espera que suas necessidades sejam atendidas ao menor custo, com um serviço adequado e com um bom atendimento. Portanto, é crucial que aspectos como as expectativas e as percepções dos

---

[55] HOFMAN, P. *et al*. Specification for resource description methods Part 2: Selection criteria for quality controlled information gateways. In: **Project RE 1004 (RE): DESIRE – Development of a european service for information on research and education**. Deliverable D3.22, mar. 1996, 90p. Disponível em: http://www.ukoln.ac.uk/metadata/desire/quality/. Acesso em: nov. 2002. p. 34

[56] Clark *et al. apud* HOFMAN, P. *et al*. *op. cit*. p. 34.

[57] Bergman e Klefsjö *apud* HOFMAN, P. *et al*. *op. cit*. p. 34.

clientes/usuários sejam levadas em conta na definição de um modelo de qualidade. Hofman e colaboradores[58]apresentam dois modelos de qualidade de serviços, o modelo Grönroos e o modelo de lacunas (gap model).

Modelo de qualidade de serviços de Grönroos

Este modelo procura entender como a qualidade de um determinado serviço é percebida pelos usuários. Para tal, a percepção do usuário é estabelecida em duas dimensões. Na primeira, a qualidade técnica, procura-se entender O QUE o consumidor recebe, ou seja, o resultado técnico do processo. Na segunda dimensão, a qualidade funcional, procura-se saber COMO o usuário daquele serviço recebe o resultado técnico ou o "desempenho significativo de um serviço" nas palavras de Grönroos.

Para Grönroos[59], no âmbito dos serviços, a qualidade funcional é percebida como sendo mais importante que a qualidade técnica, assumindo-se que o serviço foi prestado a um nível tecnicamente satisfatório. O modelo de Grönroos ressalta a importância de se incluir entre os critérios que avaliam a qualidade dos serviços, o modo como estes serviços são prestados.

Modelo de lacunas (gap model)

No modelo de lacunas procura-se identificar ou descrever as insatisfações dos usuários no contexto da qualidade do serviço. Em um estudo realizado em 1985 por Parasuraman[60] com executivos de empresas norte-americanas foram identificadas cinco "lacunas" com respeito à qualidade de serviços.

"Um conjunto de discrepâncias-chave ou "lacunas" ocorrem com respeito às percepções da qualidade de serviços e com as atividades associadas à entrega dos serviços aos usuários. Estas "lacunas" podem ser os principais obstáculos na tentativa de se prestar um serviço o qual o usuário perceberia como sendo de alta qualidade".[61]

---

[58] HOFMAN, P. *et al*. *op. cit*. p. 35.
[59] Grönroos *apud* HOFMAN, P. *et al*. *op. cit*. p. 35.
[60] Parasuraman *et al. apud* HOFMAN, P. *et al*. *op. cit*. p. 35.
[61] Parasuraman *et al. apud* HOFMAN, P. *et al*. *op. cit*. p. 35.

As cinco "lacunas" são as seguintes:

1. *Entre a expectativa do usuário e as percepções do gerenciamento destas expectativas, ou seja, o não-conhecimento do que os usuários esperam.*
2. *Entre as percepções do gerenciamento das expectativas dos usuários e as especificações de qualidade do serviço, ou seja, padrões de qualidade de serviço errados.*
3. *Entre as especificações de qualidade do serviço e a prestação do serviço, ou seja, a "lacuna" do desempenho do serviço.*
4. *Entre a prestação do serviço e a comunicação externa aos usuários sobre a prestação do serviço, ou seja, quando promessas não correspondem à prestação do serviço.*
5. *Entre a expectativa do usuário e o serviço percebido por ele (o total das quatro outras "lacunas").*

Esta última "lacuna" é a mais importante porque mostra que este modelo tem o foco voltado para a percepção do usuário.[62]

Como parte desta pesquisa Zeithman e colaboradores definiram um conjunto de dez categorias de requisitos de qualidade que ele denominou "Determinantes da qualidade de serviços". São os seguintes:

- *Tangíveis – a aparência das instalações físicas, equipamentos, pessoal e material de divulgação.*
- *Confiabilidade – habilidade para desempenhar o serviço prometido de uma maneira segura e precisa.*
- *Sensibilidade (reação) – disposição para ajudar os usuários e prover o serviço sem demora.*
- *Competência – Possuir as habilidades e o conhecimento para desempenhar o serviço.*
- *Cortesia – polidez, respeito, consideração e ser amigável no contato pessoal.*

[62] HOFMAN, P. *et al*. *op. cit*. p. 36.

- *Credibilidade – fidelidade, acredibilidade, honestidade do provedor do serviço.*
- *Segurança – livre de perigo, risco ou dúvidas.*
- *Acesso – acessibilidade e facilidade de contato.*
- *Comunicação – manter os usuários informados utilizando uma linguagem acessível.*
- *Entendendo o usuário – o esforço para conhecer os usuários e suas necessidades.* [63]

Baseado nesses dez determinantes foi desenvolvida uma escala de medida das percepções do usuário denominada SERVQUAL. Esta escala tem sido objeto de críticas e de refinamentos. Existe um debate contínuo sobre a avaliação da qualidade dos serviços e os determinantes que devem ser utilizados.[64]

As organizações estão sendo, cada vez mais, compelidas a priorizar ou dar ênfase aos seus programas de qualidade de serviços num processo de melhoria contínua. Schlesinger e Heskett[65] argumentam que as organizações deveriam abandonar os modelos adotados na indústria – técnicas de produção em massa usadas em supermercados, restaurantes "*fast food*" e aeroportos – e adotar um "novo modelo" de serviço baseado nos requisitos do usuário.

Tom Peters introduziu o conceito de "excelência" e outros conceitos como orientação ao mercado.

Um produto é o resultado de um processo organizacional podendo ser um bem (tangível) ou um serviço (intangível). O desenvolvimento do conceito de qualidade na indústria criou a necessidade de uma estrutura organizacional que pudesse incluir os conceitos de qualidade em cada estágio, desde o planejamento até a entrega do produto. Este processo foi chamado de Qualidade Total (*Total Quality Management* - TQM).

Qualidade Total ou Gestão da Qualidade significa um modo de organização com o objetivo de garantir produtos de qualidade, buscando a satisfação das pessoas

---

[63] Zeithman *et al apud* HOFMAN, P. *et al*. *op. cit*. p. 36.
[64] Parasuraman *et al apud* HOFMAN, P. *et al*. *op. cit*. p. 36.
[65] Schlesinger e Heskett *apud* HOFMAN, P. *et al*. *op. cit*. p. 36.

envolvidas em toda a cadeia do processo produtivo, sejam eles colaboradores, fornecedores, acionistas ou clientes. Trata-se de uma filosofia administrativa que visa agregar valor ao produto.[66]

A essência do TQM, segundo Bergman e Klefsjö, está baseada nos seguintes aspectos:

- *foco no cliente*
- *decisão baseada em fatos*
- *foco no processo*
- *melhoria contínua*
- *comprometimento* [67]

"*Benchmarking*" é uma outra abordagem recente que tem como objetivo assegurar uma melhoria constante na qualidade dos processos organizacionais através de um processo contínuo de comparação de produtos, serviços e práticas com líderes. A identificação e a incorporação das melhores práticas irá possibilitar às organizações um nível de desempenho elevado e sustentável.

Estes aspectos são especialmente importantes para a indústria de serviços porque enfatiza a qualidade como um processo contínuo uma vez que as percepções dos consumidores estão em constante mudança. A qualidade torna-se um processo de contínuo *feedback* e melhoria. Este conjunto de processos é conhecido como "sistema da qualidade".[68]

3.1.1. Sistema da Qualidade

O Sistema da Qualidade é definido pela norma NBR ISSO 9004-1/1994 que orienta a formulação dos procedimentos, processos e recursos necessários para implementar a gestão da qualidade.

---

[66] MOURA, L. R. Informação: a essência da qualidade. **Ciência da Informação**, Brasília, v. 25, n. 1, 1995. p.2.
[67] Bergman e Klefsjö *apud* HOFMAN, P. *et al*. *op. cit*. p. 37.
[68] HOFMAN, P. *et al*. *op. cit*. p. 37.

O sistema da qualidade tem como função assegurar as condições que garantam as especificações de qualidade dos produtos no nível operacional da organização.[69]

Um sistema da qualidade é, basicamente, um conjunto organizado de documentos que definem procedimentos, planos, registros de fatos ocorridos e responsabilidades. Esta documentação é organizada em quatro níveis, a saber:

1. Manual da Qualidade
2. Procedimentos
3. Instruções
4. Registros

O controle documental é um dos principais alicerces do sistema da qualidade. Através dele é possível assegurar o cumprimento dos requisitos estabelecidos. O acesso às informações atualizadas sobre o desenvolvimento das atividades está disponível aos colaboradores, propiciando um ambiente de melhoria contínua.[70]

3.1.2. Serviços de Informação

Os serviços de informação são estruturas organizacionais com a missão de suprir as necessidades de conhecimento requeridas pelas organizações. São os fornecedores do insumo informação.

Os serviços de informação podem ser unidades administrativas suprindo a organização com informações de interesse geral ou podem ser estruturas organizacionais com objetivos mais específicos como, por exemplo, suprir informações a uma determinada área ou setor da economia ou do governo. Núcleos de Informação Tecnológica coordenados pelo IBICT (Instituto Brasileiro de Informação Científica e Tecnológica), a rede SEBRAE, o Programa PROSSIGA e a Plataforma Lattes são exemplos de serviços de informação.

A qualidade afeta os serviços de informação de duas maneiras. Na primeira, como o serviço de informação é uma estrutura que deve atuar de maneira adequada no

[69] MOURA, L. R. *op. cit*. p. 6.
[70] VALSS, V. M. O gerenciamento dos documentos do sistema da qualidade. **Ciência da Informação**, Brasília, v. 25, n. 2, 1995.

atendimento das necessidades dos seus usuários, ele deve implementar a gestão da qualidade nos seus processos de modo melhor atender seus usuários. Na segunda maneira, o serviço de informação é afetado no sentido de que ele deve estar preparado para oferecer informações sobre qualidade, assunto cada vez mais solicitado pelas empresas e instituições. Isto é, o serviço de informação deve se capacitar nos assuntos da qualidade, seja através da busca de fontes de informações e profissionais qualificados ou através do estabelecimento de parcerias com empresas atuantes no setor.[71]

3.1.3. Sistemas de informação automatizados

Não é incomum na literatura que o assunto sistemas de informação seja iniciado com relatos sobre o elevado índice de fracasso na implementação de projetos de sistemas de informação automatizados. Relatos de sistemas mal sucedidos são apresentados a cada ano com o objetivo de se tentar descobrir as causas do fracasso e apresentar soluções para o problema. Estudos realizados no Reino Unido mostraram, por exemplo, que até 20% dos investimentos em desenvolvimento de sistemas são desperdiçados em sistemas (na forma de *softwares*) nunca entregues ou entregues mas não usados[72]. Nesses casos, a causa principal é o não atendimento de todos os requisitos dos usuários. Trata-se, portanto, de uma questão típica de falta de qualidade. Outro sério problema no desenvolvimento de sistemas de informação é a questão da produtividade. Flynn, citando estatísticas de uma empresa de desenvolvimento de *softwares* do Reino Unido revela que *30% dos maiores projetos ultrapassaram em muito seus orçamentos e cronogramas iniciais, e, quando completados, não realizaram as tarefas para os quais foram projetados*.[73] Os Estados Unidos apresentam estudos com dados semelhantes e pode-se supor que no Brasil a situação não será muito diferente.

Flynn define que um sistema bem sucedido é aquele que satisfaz seus objetivos de qualidade e produtividade. Os problemas relativos à qualidade podem ser categorizados da seguinte forma:

---

[71] MOURA, L. R. *op. cit*. p. 8.
[72] Flynn *apud* FURNIVAL, A. C. A participação dos usuários no desenvolvimento de sistemas de informação. **Ciência da Informação**, Brasília, v. 25, n. 2, p. 1-13, 1995. p.3.
[73] *Id*.

- *Enfoque errado. São escolhidas atividades erradas para se automatizar, o problema não foi definido corretamente ou pode entrar em conflito com as metas e estratégias da organização.*
- *Negligência da organização. Fatores psicológicos e sociais mais amplos podem ser negligenciados como o grau de descentralização ou centralização da organização ou o grau de aceitação ou usabilidade do sistema.*
- *Análise incorreta. As atividades corretas são identificadas mas pode-se cometer erros na análise das necessidades de informação devido a técnicas fracas de desenvolvimento.*
- *Motivos errados. Tecnocratas ou fãs das novas tecnologias com influência na organização querem implementá-las (o chamado "technology push") ou gerentes que querem estender seu poder e influência por meio do sistema computadorizado (o chamado "political pull").*[74]

O insucesso na implementação de sistemas de informação automatizados é freqüentemente atribuído a falhas nas metodologias tradicionais de análise de sistemas. Nestas metodologias muita atenção é dedicada à produção de especificações rígidas cuidadosamente documentadas. Supõe-se que o "problema" (aquilo que o novo sistema irá resolver) possa ser expresso numa base lógica, descrito em uma linguagem formal e precisa. Estas metodologias têm sua origem histórica no contexto de grandes projetos governamentais os quais, para atender processos licitatórios, exigiam especificações escritas e pormenorizadas a partir de um mesmo conjunto de requisitos gerando orçamentos e cronogramas que pudessem ser comparados na disputa para ganhar o projeto.

Kensing e Munk-Madsen observam nesse contexto da comunicação escrita, que o processo de *design* está baseado num modelo onde a realidade externa é interpretada na mente do analista de sistemas e daí transportada até aos receptores (usuários).[75] Em outras palavras, neste modelo, o receptor tem um papel passivo e a comunicação entre

[74] FURNIVAL, A. C. *op. cit*. p. 4.
[75] Kensing e Munk-Madsen *apud* FURNIVAL, A. C. *op. cit*. p. 5.

emissor-receptor será bem sucedida na medida em que o emissor (analista) esteja capacitado a formular uma mensagem rigorosa e completa.

Com esse tipo de abordagem, os críticos das metodologias tradicionais chamam a atenção sobre o fato de que são excluídos do processo de *design* fatores sociais e psicológicos da organização na qual o novo sistema será implantado. As metodologias tradicionais ignoram "os fatores humanos" dos sistemas. Como conseqüência verifica-se a resistência dos usuários manifestada através do sub-uso, do boicote e até mesmo da sabotagem do novo sistema.[76]

3.1.4. Os novos usuários dos sistemas de informação

As novas alternativas metodológicas para o projeto de sistemas de informação apresentam grande afinidade com os conceitos da qualidade, ou seja, levam em consideração a satisfação das necessidades dos usuários. Furnival observa que o argumento principal dos críticos das metodologias tradicionais era de que *o grau de usabilidade de um sistema dependia do grau de integração dos usuários ao próprio processo de design do sistema*.[77]

Nas décadas de 60 e 70, as metodologias tradicionais eram relativamente eficientes pois o processamento era feito em lote (*batch*) e, principalmente, os usuários eram profissionais de informática ou engenharia. A linguagem entre estes profissionais e o analista de sistemas eram muito próximas, o que facilitava a comunicação entre eles.

Nos dias atuais, o perfil do usuário modificou-se totalmente. Profissionais de todas as áreas, como, por exemplo, cientistas, artistas, burocratas, advogados, os chamados usuários finais interagem hoje com sistemas *on-line* para atender aos mais diversos tipos de necessidades informacionais (profissionais, culturais, lazer, bancárias, compras, etc). A exposição a esta variedade de sistemas aplicativos automatizados torna os usuários mais exigentes, seletivos e críticos.

Portanto, para se adequarem aos novos requisitos de qualidade, as metodologias tradicionais foram obrigadas a incorporar estas novas características, destacando-se

[76] FURNIVAL, A. C. *op. cit*. p. 5.
[77] FURNIVAL, A. C. *op. cit*. p. 5.

aquelas relativas à participação dos usuários no processo de desenvolvimento dos sistemas. São as chamadas metodologias de "*participatory design*".[78]

Segundo Maturana e Varela, *a comunicação depende não do que é transmitido mas do que acontece à pessoa que recebe*.[79] Este enunciado sustenta a maioria das metodologias de *design* participativo. Em outras palavras, os usuários (receptores) participam da atividade de comunicação com o analista (emissor). Do ponto de vista da qualidade, à medida que os usuários colaboram com os analistas por meio destes contatos, estes analistas estão absorvendo o que é visto como "o necessário" do domínio dos usuários.

Segundo Booth *o objetivo final do design participativo é tornar melhor a qualidade de vida dos profissionais na organização por meio do enriquecimento do seu trabalho, usando a tecnologia para contribuir na realização deste objetivo, e não usando-a por usá-la*. [80]

3.2. Qualidade aplicada às bases de dados

Historicamente, estudos e pesquisas em qualidade de bases de dados adquirem importância no final dos anos 80 e início dos anos 90.[81] No início da utilização das bases de dados qualquer resultado era uma grande conquista pela velocidade na sua obtenção em comparação aos lentos sistemas manuais, ficando a questão da qualidade do dado em segundo plano. Com a rápida vulgarização e disseminação dos sistemas *on-line* as bases de dados evoluíram rapidamente, em particular, no tocante às bases bibliográficas. No início, eram utilizadas como sistemas de recuperação, passando, posteriormente, para bases de dados de texto completo.[82]

No âmbito das bibliotecas, qualidade da informação não era um aspecto muito considerado antes do aparecimento das bases de dados eletrônicas. As bibliotecas

---

[78] FURNIVAL, A. C. *op. cit*. p. 5.
[79] Maturana e Varela *apud* FURNIVAL, A. C. *op. cit*. p. 6.
[80] Booth *apud* FURNIVAL, A. C. *op. cit*. p. 9.
[81] HEEMANN, V. **Avaliação ergonômica de interfaces de bases de dados por meio de *checklist* especializado**. Orientador: Walter de Abreu Cybis. Florianópolis: UFSC, 1997. Dissertação. (Mestrado em Engenharia da Produção). Disponível em: http://www.eps.ufsc.br/disserta97/heemann/. Acesso em: nov. 2002.
[82] HEEMANN, V. *op. cit*.

selecionavam os livros e revistas de acordo com seus próprios critérios os quais normalmente atendiam às necessidades dos seus usuários ou da organização. O foco em qualidade da informação só despertou um interesse maior com o uso crescente das bases de dados eletrônicas, tanto *on-line* como em CD-ROM.[83]

Com o advento da Internet surgiram novos paradigmas exigindo novas abordagens para a disponibilização de grandes bases de dados. Inicia-se o surgimento de estudos relacionados aos aspectos de acesso e utilização visando estabelecer critérios mínimos para o oferecimento dessas bases de dados de maneira eficiente nesses novos ambientes.[84]

Nesse contexto, surgem iniciativas de mobilização que merecem destaque: o *Centre for Information Quality Management* – CIQM e o *Southern California User Group* – SCOUG.

O CIQM foi criado pelos *The Library Association* e o *UK Online User Group* para atuar como um fórum de discussão. Neste fórum os usuários relatam seus problemas ligados à qualidade das bases de que fazem uso e o CIQM se encarrega de encaminhar o problema ao provedor da informação e de, posteriormente, fazer retornar ao usuário uma resposta. Este serviço é gratuito para os usuários.[85]

O grupo SCOUG foi outra importante iniciativa. Em 1990, juntamente com a *British Library Research and Development Department* – BLR & DD criaram uma lista de critérios de qualidade para o uso de bases de dados. Mais adiante estes critérios serão vistos em detalhe.

Heemann[86] chama atenção para um problema cada vez mais freqüente que é a descentralização da alimentação das bases de dados em redes. A base Currículo Lattes é um bom exemplo dessa tendência de descentralizar megabases. Armstrong aponta a

---

[83] HOFMAN, P. *et al*. *op. cit*. p. 38.
[84] HEEMANN, V. *op. cit*.
[85] ARMSTRONG, C. Metadata, PICS and quality. **Ariadne**, v. 9, maio 1997. Disponível em: http://www.ariadne.ac.uk/issue9/pics/. Acesso em: dez. 2002.
[86] HEEMANN, V. *op. cit*.

falta de crítica dos sistemas e acrescenta que, nesses ambientes, *os dados supridos por terceiros, individuais ou institucionais, são assumidos como corretos nos sistemas.* [87]

Armstrong[88] acredita que, em geral, os usuários tendem a julgar uma nova base de dados através das informações apresentadas nos catálogos dos fornecedores. Verifica-se muitas vezes que pesquisas realizadas podem exceder a capacidade das bases de dados. Por exemplo, raros sãos os usuários que conhecem a política do provedor da informação com respeito à inclusão de dados: algumas bases indexam todos os artigos de um periódico, outras, apenas os artigos-chave, outras, ainda, podem variar as regras em função do periódico.

Twidale e Marty[89], em recente revisão bibliográfica descrita em artigo sobre qualidade de dados, observam que o tema tem sido de interesse de pesquisadores de diversas áreas como Informática, Biblioteconomia, Ciência da Informação e Sistemas de Informação Gerenciais. Com relação a esta última área existe um forte interesse comercial voltado principalmente para as questões dos custos em organizações comerciais com dados de baixa qualidade.[90]

Medawar[91] fez uma revisão da literatura em qualidade de bases de dados no contexto da Ciência da Informação, abordando suas relações com a TQM (*Total Quality Management*) e o foco na satisfação do usuário.

Ballou e Tayi[92] relatam a importância de estabelecer prioridades no esforço de se obter a melhoria da qualidade dos dados e propõe modelos para determinar estas prioridades baseadas em análise de custo-benefício.

Jasco[93] avalia aspectos da qualidade de dados a partir da perspectiva do usuário final de uma base de dados. Ele observa a quantidade espantosa de "lixo" nas bases de dados

---

[87] Armstrong *apud* HEEMANN, V. *op. cit*.
[88] ARMSTRONG, C. *op. cit*.
[89] TWIDALE, M. B.; MARTY, P. F. An investigation of data quality and collaboration. **Technical Report ISRN UIUCLIS--1999/9+CSCW**, 1999. Disponível em: http://www.lis.uiuc.edu/~twidale/pubs/dq.html. Acesso em: dez. 2002.
[90] Redmond *apud* TWIDALE, M. B.; MARTY, P. F. *op. cit*.
[91] Medwar *apud* TWIDALE, M. B.; MARTY, P. F. *op. cit*.
[92] Ballou e Tayi *apud* TWIDALE, M. B.; MARTY, P. F. *op. cit*.
[93] Jasco *apud* TWIDALE, M. B.; MARTY, P. F. *op. cit*.

numéricas e bibliográficas, em especial, valores ausentes em certos campos e como isto pode levar a resultados enganosos.

Em outro artigo de 1993, Jasco constata a contribuição que o uso efetivo de vocabulários controlados pode trazer para o problema da qualidade de dados.

Problemas permanecem com as bases de dados comerciais, incluindo erros ortográficos e de digitação e o uso de alguns campos de dados como uma "área de despejo" (*dumping ground*) para valores que não se encaixam na estrutura de campos da base de dados corrente. Problemas posteriores são causados pelas variantes ortográficas legítimas, especialmente nos casos onde os nomes mudam com o tempo. Nesses casos, o uso de referências cruzadas pode ser uma solução eficaz.

Wang e colaboradores[94] exploram a especificação de metadados com o foco voltado para a qualidade dos dados. Neste estudo observou-se que seria insuficiente ter apenas uma única medida de qualidade para um registro. Cada elemento de um registro poderia apresentar informações distintas sobre qualidade. Uma abordagem semelhante é proposta por Armstrong com o uso de rótulos de qualidade nas bases de dados (Database Labels).[95]

Seguindo a linha proposta por Twindale e Marty na qual se propõe o uso do *feedback* do usuário final, Davis (1989) descreve um trabalho realizado pela OCLC (*Online Computer Library Center*) no qual foram trazidas à tona as impressões dos seus usuários sobre a qualidade da sua base *Online Union Catalog*. O processo foi trabalhoso com os usuários (bibliotecários) envolvendo o envio de formulários e documentação pelo correio. Nesse estudo foi constatado que 31% dos respondentes disseram que nunca relatam erros e 42% somente relatam alguns erros. O que chamou atenção foi o fato de existirem alguns usuários que não se incomodam em relatar erros. Tal constatação mostrou que a metodologia do "*feedback*" é viável, reforçada pelo fato de que 70% dos bibliotecários entrevistados disseram que aumentariam seus relatórios de erros caso fosse colocado à disposição deles um sistema "*on-line*" mais acessível.[96]

---

[94] Wang *et al. apud* TWIDALE, M. B.; MARTY, P. F. *op. cit.*
[95] ARMSTRONG, C. Metadata, PICS and quality. **Ariadne**, v. 9, maio 1997. Disponível em: http://www.ariadne.ac.uk/issue9/pics/. Acesso em: dez. 2002.
[96] TWIDALE, M. B.; MARTY, P. F. *op. cit.*

Orr também considera o "*feedback*" do usuário uma alternativa metodológica para a melhoria da qualidade dos dados. Havendo um sistema de controle do "*feedback*", existe a possibilidade do uso de estatística para a detecção e correção de erros. Esta análise teve como resultado a definição de seis regras da qualidade de dados, reproduzidas a seguir:

- *Dados não utilizados não podem permanecer corretos por muito tempo.*
- *Qualidade de dados em um sistema de informação é uma função do seu uso, não da sua coleção.*
- *Qualidade de dados, em última análise, não será melhor do que seu uso mais estrito.*
- *Problemas de qualidade de dados tendem a tornar-se piores com o envelhecimento do sistema.*
- *Quanto menos provável algum atributo de dado (elemento) está para mudar, mais traumático será quando ele, finalmente, sofrer a mudança.*
- *Leis da qualidade de dados se aplicam igualmente a dados e a metadados.* [97]

Bowen e colaboradores[98] discutem maneiras pelas quais as organizações usam técnicas estatísticas para obter a melhoria contínua na qualidade de bases de dados "persistentes" (p. ex.: ativos fixos, inventários, informações ao consumidor, bibliografias, etc.). Tais técnicas estatísticas são aplicadas de diversas maneiras. Por exemplo, estabelecendo relações entre a qualidade dos dados e a vida útil de um conjunto de dados. O estudo descreve as relações entre o gerenciamento dos processos estatísticos e o processamento das transações. Mostra como as organizações podem se antecipar e se prevenir de problemas com dados e melhorar continuamente a qualidade dos mesmos. Os autores acreditam que a implementação dessas estratégias pode ajudar aos gerentes a desenvolver uma cultura de melhoria da qualidade dos dados.

---

[97] Orr *apud* TWIDALE, M. B.; MARTY, P. F. *op. cit*.

[98] BOWEN, P. L.; FUHER, D. A. GUESS, F. M. Continuously improving data quality in persistent databases. **Data Quality**, Alexandria, EUA, v. 4, n. 1, set. 1998. Disponível em: http://www.dataquality.com/998bowen.htm. Acesso em: dez. 2002.

Drucker[99] afirma que a melhoria da qualidade dos dados é um tema cada vez mais importante pelo fato de que os sistemas de informação estão se tornando cruciais para que as organizações possam explorar as oportunidades que a Ciência da Informação e as tecnologias da informação irão proporcionar nesta década.

Governos e organizações cada vez mais contam com seus sistemas de informação para integrar e dar suporte aos seus processos de tomada de decisão. Esses sistemas e a qualidade dos dados neles contidos afetam a percepção dos usuários na qualidade dos produtos e serviços adquiridos.[100] Dados imprecisos reduzem o valor dos sistemas de informação e levam a decisões pobres. Pesquisas em economia da informação demonstram que a precisão é o determinante mais importante no valor de um sistema de informação.[101]

O interesse em qualidade de dados pode ser observado pelos esforços em modelar, melhorar e definir este conceito. Pesquisadores desenvolveram modelos quantitativos para ajudar auditores, controladores e desenvolvedores de sistemas a avaliar, melhorar e gerenciar a precisão dos dados. Alguns desses modelos foram construídos a partir dos dados acumulados nos próprios sistemas de informação. Outras abordagens defendem o uso de restrições de integridade baseadas em regras, análise de integridade, gestão da qualidade total de dados e procedimentos de controle a nível de tabela para melhorar a qualidade dos dados entrantes e acumulados.[102]

Hernandez-Orallo analisa aspectos envolvendo técnicas de KDD (*Knowledge Discovery in Databases*) e qualidade de dados. KDD é definido como um *processo não-trivial para identificar dados válidos, novos, potencialmente úteis, e, em última análise, padrões compreensíveis nos dados*.[103] A preparação e, em especial, a limpeza dos dados é o aspecto mais crítico em KDD. A precisão dos dados tem sido considerada como o maior problema em muitos sistemas dinâmicos e a principal razão para esta perda de precisão é o tempo. À medida que o tempo passa, mais e mais as informações se tornam obsoletas. Ele considera que os métodos tradicionais resolvem em parte este problema e

---

[99] Drucker *apud* BOWEN, P. L.; FUHER, D. A. GUESS, F. M. *op. cit.*
[100] Wang e Strong *apud* BOWEN, P. L.; FUHER, D. A. GUESS, F. M. *op. cit.*
[101] Hilton *et al. apud* BOWEN, P. L.; FUHER, D. A. GUESS, F. M. *op. cit.*
[102] BOWEN, P. L.; FUHER, D. A. GUESS, F. M. *op. cit.*
[103] HERNÁNDEZ-ORALLO, J. **Knowledge discovery in databases and data quality**. 1999. Disponível em: http://www.dsic.upv.es/~jorallo/KDD/KDD.html. Acesso em: dez. 2002.

propõe que, ao invés do uso de técnicas de medição da qualidade "*a posteriori*", seria mais razoável a realização de medições contínuas da satisfação e da interação do usuário em relação à informação que está armazenada na base de dados. A exatidão dos dados contidos na base seria obtida pelas freqüências de correção em vez da precisão dos "*experts*".

Algumas técnicas de medição da qualidade seriam baseadas na auditagem de amostras de dados, comparando-se uma visão parcial da realidade com uma visão parcial da base de dados. Segundo Hernández-Orallo o problema principal desta comparação é a de que não existiria um modo fácil de saber qual é a realidade verdadeira, porque isto dependeria em muito da precisão do "*expert*" ou do usuário que interpreta a realidade.[104]

Pedrini descreve um método para avaliar a fidedignidade de referências bibliográficas registradas em bases de dados. A técnica utilizada consistia em avaliar cada elemento descritivo da referência, utilizando-se como indicadores para medir a fidedignidade das referências os conceitos de completude, correção e normalização (norma NBR-6023).[105]

Wang e Strong acreditam que os usuários de dados têm uma concepção de qualidade de dados que vai além da precisão dos dados, foco da maioria dos esforços empreendidos pelos pesquisadores e empresas. Neste sentido, eles definiram uma estrutura que captura os aspectos da qualidade de dados que são importantes para os usuários. Ressaltam que a característica desse estudo em particular é a de que os atributos de qualidade dos dados são obtidos dos usuários ao invés de terem sido definidos teoricamente ou baseado nas experiências dos pesquisadores. Consideram que *dados de alta qualidade devem ser intrinsecamente bons, contextualmente apropriados para a tarefa, claramente representados e acessíveis ao usuário dos dados.*[106]

---

[104] HERNÁNDEZ-ORALLO, J. *op. cit*.

[105] PEDRINI, A. G. **O cientista e os métodos de avaliação de seu desempenho: estudo de sua adequação no contexto brasileiro**. Orientador: Rosali Fernandez de Souza. Rio de Janeiro: UFRJ, Escola de Comunicação – CNPq/IBICT, 1999. 442p. Tese. (Doutorado em Ciência da Informação).

[106] WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information Systems**, v. 12, n. 4, p. 5-33, 1996.

3.3. Critérios de qualidade para bases de dados

As primeiras bases de dados *on-line* eram tratadas com muita "reverência" por seus usuários. Por esta razão, é que somente no final dos anos 80 os usuários de bases de dados começaram a dar sugestões no sentido de melhorá-las. Neste sentido, como já mencionado anteriormente, uma das mais importantes iniciativas de usuários de bases de dados foi a criação do grupo SCOUG - *Southern California Online Users Group*. O grupo SCOUG é formado por usuários do meio acadêmico, bibliotecas universitárias e de corporações e instituições de pesquisa. É uma organização sem fins lucrativos, *dedicada a ajudar as pessoas a tirar o melhor proveito da informação disponível em bancos de dados on-line, na Internet, e em outros formatos eletrônicos*.[107]

Em 1989, o grupo SCOUG desenvolveu uma "lista de desejos dos usuários" e, no ano seguinte, apresentou uma lista de critérios de qualidade para bases de dados.[108]

A lista foi organizada em um conjunto de dez categorias:

1. Consistência.
2. Cobertura e escopo.
3. Oportunidade.
4. Taxa de erro/precisão.
5. Facilidade de uso.
6. Integração.
7. Saídas.
8. Documentação.
9. Suporte e treinamento do usuário.
10. Razão custo/benefício.

Nos anos seguintes outros autores também apresentaram listas de critérios de qualidade para bases de dados. De uma maneira geral estes critérios eram coincidentes com aqueles propostos pelo SCOUG. O modelo do SCOUG, apesar de produzido há mais de uma década, continua sendo referenciado na literatura especializada.[109]

---

[107] http://www.scougweb.org
[108] Basch *apud* HOFMAN, P. *et al*. *op. cit*. p. 38.
[109] PEREIRA, M. N. F. *op. cit*.

Em 1994, o modelo do SCOUG foi utilizado em uma pesquisa de opinião levada a cabo em doze países europeus e teve como objetivo conhecer os dez mais importantes critérios de qualidade para as bases de dados. O resultado desta pesquisa, em ordem decrescente de importância, foi o seguinte:

1. Cobertura.
2. Acessibilidade.
3. Atualidade.
4. Consistência
5. Precisão.
6. Valor.
7. Documentação.
8. Harmonização.
9. Saídas.
10. Suporte.[110]

3.3.1. Projeto DESIRE

A proximidade dos temas bases de dados *on-line* e serviços de informação disponíveis através da Internet justificam o destaque para o trabalho apresentado a seguir. Trata-se de um estudo realizado em 1996 pelo UKOLN (*UK Office for Library and Information Networking)* no âmbito de um projeto denominado DESIRE – *Development of a European Service for Information on Research and Education*. O UKOLN é um centro de especialistas em gestão da informação digital situado no campus da Universidade de Bath (Inglaterra).

O estudo[111], entitulado "*Selection Criteria for Quality Controlled Information Gateways*", teve como objetivo principal definir critérios de qualidade para serem utilizados na seleção de recursos informacionais para os portais temáticos do DESIRE. Entende-se como portal temático (*subject gateway*) um *site* na Internet que organiza e

---

[110] PEREIRA, M. N. F. *op. cit*.
[111] HOFMAN, P. *et al*. Specification for resource description methods Part 2: Selection criteria for quality controlled information gateways. In: **Project RE 1004 (RE): DESIRE – Development of a european service for information on research and education**. Deliverable D3.22, mar. 1996, 90p. Disponível em: http://www.ukoln.ac.uk/metadata/desire/quality/. Acesso em: nov. 2002.

disponibiliza acessos a diferentes recursos como, por exemplo, bases de dados *on-line* relacionadas a um tema específico. O Portal da CAPES (www.periodicos.capes.gov.br) e a Plataforma Lattes (lattes.cnpq.br) são dois exemplos de portais de informação científica e tecnológica.

O estudo do UKOLN descreve métodos e ferramentas criadas para ajudar os profissionais dos portais temáticos a desenvolver e manter seus sistemas de controle de qualidade. Destaca-se a criação de duas ferramentas. A primeira é um modelo conceitual do funcionamento de um portal temático que permitiria uma abordagem sistemática dos aspectos da qualidade no desenvolvimento, controle, monitoração e análise de um portal. O modelo foi desenvolvido com a característica de ser genérico, não sendo restrito a qualquer área temática em particular. Ele teria a capacidade de identificar pontos-chave nos quais os critérios de qualidade poderiam ser empregados.[112]

A segunda ferramenta, de maior interesse para o presente trabalho, baseava-se em uma lista de critérios de qualidade para ser empregada na seleção de recursos informacionais. Esta lista estruturada de critérios poderia ser utilizada tanto como uma ferramenta de referência pelos portais existentes como também permitiria que novos portais pudessem produzir seus próprios esquemas de seleção.[113]

A criação da lista levou em conta aspectos gerais que envolvem o processo de seleção de um recurso para um determinado portal, tais como: os usuários, os recursos de informação e o serviço em si mesmo. Dessa análise resultou um conjunto de critérios subdivididos em cinco categorias principais de critérios de qualidade de seleção:[114]

1. Critérios de escopo: considerando os usuários.
2. Critérios de conteúdo: avaliando a informação.
3. Critérios de forma: avaliando o meio.
4. Critérios de processo: avaliando o sistema.
5. Critérios de gerenciamento da coleção: considerando o serviço.

---

[112] HOFMAN, P. *et al*. *op. cit*. p. 6.
[113] HOFMAN, P. *et al*. *op. cit*. p. 6.
[114] HOFMAN, P. *et al*. *op. cit*. p. 15.

Nesse modelo de avaliação, um "recurso de qualidade" é definido tendo sempre em mente o serviço específico e seus usuários. A partir de cada um dos cinco subconjuntos, os critérios mais adequados ao serviço específico devem ser selecionados e continuamente revisados.[115]

Os critérios definidos pelo estudo do projeto DESIRE, organizados de acordo com as cinco categorias acima mencionadas são apresentados em detalhe a seguir.

Critérios de escopo: considerando os usuários

- cobertura da informação
- acesso
- políticas de catalogação
- aspectos geográficos

Os critérios ou as políticas relacionadas ao escopo do serviço avaliado determinam o que será ou não incluído no catálogo. Por isso, os critérios de escopo são os primeiros filtros na seleção do recurso. Tudo que fica fora do escopo será rejeitado e o que ficar dentro será submetido ao restante do processo de seleção de qualidade. Os aspectos mais importantes a serem considerados na escolha dos critérios de escopo, para um determinado serviço, são os propósitos do serviço e o público-alvo.[116]

Critérios de conteúdo: avaliando a informação

Estes critérios estão baseados mais no conteúdo informacional dos recursos e menos no fato de estarem disponibilizados na Internet. Estão relacionados aos critérios tradicionais utilizados pelas bibliotecas na seleção de livros e periódicos, como validade, autoridade e reputação das fontes, precisão, abrangência, composição, organização e originalidade das informações.[117]

---

[115] HOFMAN, P. *et al*. *op. cit*. p. 15.
[116] HOFMAN, P. *et al*. *op. cit*. p. 17.
[117] HOFMAN, P. *et al*. *op. cit*. p. 20.

Critérios de forma: avaliando o meio

Critérios de forma estão relacionados à apresentação e à organização da informação. Alguns destes critérios seriam os mesmos aplicáveis em recursos disponíveis em papel. Outros seriam definidos em função do meio eletrônico, no caso a Internet. Estes critérios fariam referência a aspectos como facilidade de "navegar" e de pesquisar o recurso informacional, suporte ao usuário, uso de padrões reconhecidos, uso apropriado da tecnologia e aspectos estéticos.[118]

Critérios de processo: avaliando o sistema

Os critérios de processo estão baseados nos processos que dão suporte ao recurso informacional. Neste caso, ao contrário dos critérios de conteúdo e forma, estes estariam intimamente relacionados ao fato de que são recursos de Internet. E, como a informação na Internet pode comprometer a integridade de um trabalho publicado, tal fato faz provocar o surgimento de inúmeras questões sobre a qualidade do recurso ao longo do tempo. Os critérios de processo estariam relacionados aos seguintes aspectos:[119]

1. Integridade da informação – responsabilidade do provedor de informação. Envolve questões sobre atualidade e freqüência de atualização da informação, adequação da manutenção das informações, etc.

2. Integridade do *site* – responsabilidade do *webmaster*. Envolve questões sobre a atualização, durabilidade e gerenciamento do *site*.

3. Integridade do sistema – responsabilidade do administrador do sistema. Envolve questões sobre o desempenho técnico do recurso, estabilidade, confiabilidade e integridade do sistema.

Critérios de gerenciamento da coleção: considerando o serviço

As políticas de gerenciamento da coleção de um serviço de informação determinam como os recursos serão relacionados ou descartados sob o ponto de vista da coleção

[118] HOFMAN, P. *et al*. *op. cit*. p. 24.
[119] HOFMAN, P. *et al*. *op. cit*. p. 26.

como um todo. Nesse contexto, o termo "coleção" se refere aos itens correntemente descritos no catálogo ou indicados pelo mesmo. Envolve questões comparativas entre recursos disponibilizados tanto dentro da própria coleção como fora dela. Por exemplo, o crescimento da coleção ao longo do tempo traz a necessidade de se reavaliar a existência de recursos em duplicata ou que já não possuem os níveis de qualidade exigidos num dado momento. O valor relativo de um recurso disponível na coleção em comparação com um recurso semelhante em outra coleção pode determinar a manutenção ou descarte deste recurso.[120]

3.4. Controle de qualidade de bases de dados

Controle de qualidade inclui técnicas, atividades e filosofia de gerenciamento necessárias à produção de um bem ou serviço de qualidade que satisfaça as necessidades de seus usuários. No tocante às bases de dados, o controle de qualidade envolve todas as etapas de manuseio da informação, da sua criação ao uso final. Além da qualidade intrínseca da base de dados, a qualidade do produto informacional é influenciada pelo *hardware*, *software* de processamento e recuperação, telecomunicação, documentação e a assistência ao usuário.

Armstrong[121] ressalta os problemas que podem resultar de erros nas bases de dados. Por exemplo, erros simples como erros tipográficos podem remover registros relevantes do resultado de uma busca, comprometendo seriamente uma pesquisa. Heemann chama atenção do fato de que um dos problemas críticos em relação à qualidade de bases de dados é o das metodologias para controlar ou monitorar essa qualidade. Armstrong resume alguns dos principais problemas que costumam afetar as bases de dados, a saber:

- Campos vazios. Problema freqüentemente detectado, podendo afetar os resultados de uma pesquisa. Se um percentual de registros não incluir, por exemplo, o tipo de documento, e for solicitada uma pesquisa com esse requisito, os resultados serão vazios, mesmo que relevantes para o usuário.
- Duplicação de registros.
- Dados incorretos.

[120] HOFMAN, P. *et al*. *op. cit*. p. 28.
[121] Armstrong *apud* HEEMANN, V. *op. cit*.

- Lacunas entre os dados, provocadas por problemas de cópia e atualização da base em diferentes suportes (on-line, CD-Rom, disquete).
- A falta de padronização ou controle de autoridade.

Uma pesquisa realizada junto a usuários de bases de dados pela *European Association of Information Services* – EUSIDISC identificou os principais problemas de qualidade que afetam as bases como sendo os seguintes:[122]

1. Registros recuperados irrelevantes.
2. Muito tempo despendido na pesquisa.
3. Necessidade de se repetir pesquisas.
4. Número insuficiente de registros recuperados.
5. Registros recuperados não necessários.

3.4.1. Métodos de controle de qualidade de bases de dados

O sucesso do controle de qualidade das bases de dados requer a combinação da aplicação de métodos manuais e automatizados. Os métodos automatizados normalmente complementam os métodos manuais e raramente os elimina.

3.4.1.1. Métodos manuais

Os métodos manuais mais importantes para o controle de qualidade de bases de dados são o treinamento, a revisão e a assistência do usuário.[123]

O treinamento do usuário é considerado como o método mais básico para se obter qualidade na entrada de dados nas bases. De maneira geral existe tanto por parte dos usuários como dos provedores de informação uma certa falta de interesse ou negligência em, respectivamente, acessar ou disponibilizar informações sobre os elementos de dados, os formatos e os padrões para entrada de dados e outros procedimentos. A documentação destas informações constitui-se na parte essencial da educação e

[122] HEEMANN, V. *op. cit*.
[123] O'NEIL, E. T.; VIZINE-GOETZ, D. Quality control in on-line databases. In: WILLIAMS, M. E., ed. **Annual review of information science and technology (ARIST)**. New Jersey: Elsevier-ASIS, v. 23, 1988. p. 125-156. p. 130.

treinamento do usuário. Problemas de baixa qualidade dos dados estão relacionados, muitas vezes, a falhas na documentação. Para que o usuário tenha a seu dispor uma ajuda efetiva, a documentação deve estar sempre atualizada e completa. Conhecendo a cultura imediatista do usuário, normalmente refratária à consulta a manuais e procedimentos, muitos provedores de informação oferecem números de telefone e endereços de correio eletrônico para dar suporte aos usuários. Mais recentemente, alguns provedores têm oferecido a possibilidade de consulta de ajuda *on-line* através do uso de *softwares* de bate-papo (*chat* - comunicação em tempo real entre duas ou mais pessoas através de texto, voz ou imagem).

Os métodos manuais de revisão, apesar de trabalhosos e caros, oferecem a oportunidade de eliminar muitos erros, especialmente, se tal revisão puder ser realizada antes da entrada de dados na base. Reeb[124] relata que na revisão de folhas de registros de produção de um catálogo foi encontrada uma média de 0,6 erros por registro. Mantendo-se estatísticas de tipos e freqüências de erros encontrados na revisão manual, é possível identificar áreas problemáticas as quais, posteriormente, poderão se objeto de atenção especial para sanar ou minimizar os problemas detectados. Para alguns tipos de bases de dados a revisão manual é uma atividade crítica. Apenas para citar como exemplo, o Banco de Dados de Toxicologia, disponibilizado pela MEDLARS, contém somente dados avaliados criticamente com respeito à qualidade e integridade por especialistas da área de saúde.[125]

Outro método manual de controle de qualidade é a assistência do usuário. Esse método é normalmente implementado nas interfaces das bases de dados através de *softwares* aplicativos específicos que têm como objetivo permitir ao usuário da base relatar erros encontrados. Possui, entretanto, a desvantagem de consumir tempo do usuário e essa percepção do usuário é real, haja vista que, de um modo geral, o usuário não se sente encorajado a participar desse tipo de atividade. Conforme já relatado anteriormente, Busch realizou uma pesquisa com 141 bibliotecas, membros da OCLC, que abordava os procedimentos de relatórios de erros adotados pelo OCLC. A pesquisa mostrou que menos de 35% das bibliotecas pesquisadas relatavam erros de forma rotineira. Por outro

[124] Reeb *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 130.
[125] Eakin e Harron *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 130.

lado, questionadas de que maneira a OCLC poderia melhorar o controle de qualidade, 79% das bibliotecas selecionaram, a partir de uma lista de opções, a implementação de relatórios de erros *on-line*.

3.4.1.2. Métodos automatizados

É importante ressaltar que os métodos automatizados não substituem os métodos manuais. Um controle de qualidade eficiente para as bases de dados exige necessariamente a utilização de múltiplas abordagens combinando métodos manuais e automatizados.

O trabalho de O'Neill e Vizine-Goetz[126] destaca os seguintes métodos automatizados: correção de erros ortográficos, validação automática de dados, dados auto-verificáveis, controle de autoridade e detecção de duplicação.

Correção de erros ortográficos

Erros de ortografia incluem também outros tipos de erros similares como erros de digitação, erros provenientes da digitalização de textos através de técnicas de OCR (reconhecimento óptico de caracteres) e erros de transmissão. Portanto, qualquer erro que resulta em uma não-palavra é considerado um erro de ortografia.

Erros de ortografia parecem ser os erros mais comuns encontrados nas bases de dados. Isto pode ser explicado pelo fato de que também os erros de ortografia são os mais fáceis de serem detectados pelos usuários das bases, enquanto que outros tipos de erros são menos óbvios de serem identificados. Damerau[127] identificou os quatro tipos de erros mais comuns: omissão, inserção, substituição e transposição. Diferentes estudos indicam que estes quatro tipos de erros respondem por 80 a 96% dos erros de ortografia encontrados nas bases de dados.

---

[126] O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 132.
[127] Damerau *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 133.

Mitton[128] identifica o erro de "palavra-real" como aquele que ocorre quando um erro ortográfico resulta em uma palavra válida. Por exemplo, escrever "filha" quando a intenção era escrever "falha". A literatura não apresenta estatísticas da freqüência desse tipo de erro em textos digitados. Muitos pesquisadores tratam os erros de "palavra-real" como sendo erros gramaticais.

Outros tipos de erros ortográficos têm sido identificados: erros de divisão de palavras (exemplo: "de baixo") ou quando duas palavras aparecem juntas (exemplo: "emcima").

Com relação aos mecanismos de detecção de erros ortográficos, a maioria das pesquisas nesse campo e suas aplicações são desenvolvidas por empresas privadas, o que torna restrito o acesso a esses mecanismos por serem tecnologias proprietárias.

Quadro 1. Tipos mais comuns de erros de ortografia

Tipo de erro	Definição	Exemplo: barco
Omissão	uma letra é omitida	baro
Inserção	uma letra é adicionada	barcro
Substituição	uma letra é substituída	borco
Transposição	troca de letras adjacentes	bacro

Fonte: baseado em O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*.

Devido ao valor comercial dos "softwares" de correção ortográfica poucos pesquisadores divulgam detalhes. Em uma pesquisa realizada por Seymour[129] em 55 processadores de texto, verificou-se que quase 90% deles incluíam algum tipo de mecanismo de detecção de erros.

O processo de correção de erros envolve duas etapas: a detecção e a identificação das possíveis correções. Atualmente, a maioria dos "softwares" de detecção de erros é interativa. O conteúdo textual do documento sofre um processo de varredura e quando um erro é identificado, o "software" produz uma lista de possíveis correções. O uso de dicionários constitui-se em uma das técnicas mais bem sucedidas na correção de erros

[128] Mitton *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 134.
[129] Seymour *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 135.

ortográficos. Um dicionário padrão contém cerca de cem mil palavras. Normalmente, utilizam-se dicionários comerciais consagrados. Entretanto, para bases de dados especializadas é necessário o desenvolvimento de dicionários específicos.[130]

O processo de utilização de dicionários compara cada palavra de um texto com as palavras do dicionário. Quando surge uma palavra que não consta no dicionário, essa palavra é identificada como um possível erro. Neste processo, o aspecto mais crítico é a definição do que seja uma palavra (em um conteúdo textual em meio eletrônico). Damerau definiu uma palavra como sendo uma cadeia de caracteres terminada por um espaço em branco, uma vírgula, um ponto, uma barra ou um parêntese. Essa definição tem sido largamente aceita podendo sofrer algumas pequenas modificações. Certas classes de palavras como nomes próprios, nomenclatura da área de química e acrônimos, produzem algumas dificuldades no uso de dicionários.

Além das técnicas baseadas em dicionários a literatura identifica outras duas importantes metodologias de detecção de erros: análise de "n-gramas" (combinações entre caracteres de uma palavra) e análise de palavras de baixa freqüência.

Todos os métodos atuais estão voltados para a correção de erros em palavras analisadas separadamente. O próximo desafio será no sentido da criação de algoritmos que façam a detecção e correção de palavras considerando estas como parte de um contexto.[131]

Controle de autoridade

Um arquivo de autoridade é um *conjunto de registros que indicam a forma correta de cada entrada estabelecida*. São como dicionários especiais que podem ser utilizados para a correção de erros ortográficos e de digitação.[132]

O controle de autoridade envolve um conjunto de processos que cobre desde a criação, gravação, manutenção dos dados de autoridade até o uso efetivo dos registros e arquivos de autoridade de forma assegurar a consistência de um determinado arquivo.[133]

---

[130] O'NEIL, E. T.; VIZINE-GOETZ, D. op. cit. p. 136.
[131] O'NEIL, E. T.; VIZINE-GOETZ, D. op. cit. p. 140.
[132] PEREIRA, Maria de Nazaré Freitas. **Por uma Economia do Conhecimento**: Avaliação de Bases de Dados Nacionais para a Produção de Indicadores de C&T (Ciência e Tecnologia). Relatório Parcial (Avaliação de qualidade de bases de dados bibliográficas). Rio de Janeiro, julho/2001. Processo 520416/93-7 (NV).

O'Neill e Vizine-Goetz observam que o desenvolvimento do formato MARC para dados de autoridade e a distribuição de registros eletrônicos pela LC *(Library of Congress*) concomitantemente ao crescente uso pelas bibliotecas de registros bibliográficos eletrônicos, estimulou bibliotecários, fornecedores e outras unidades bibliográficas a desenvolverem sistemas automatizados de controle de autoridade.

Burger[134] classifica sistemas *on-line* de autoridade em três categorias: 1. Sistemas com arquivos de autoridade completamente independentes e separados das bases de dados. 2. Sistemas com arquivos de autoridade estreitamente relacionados à base de dados mas sem estar ligado a ela. 3. Sistemas com arquivos de autoridade ligados à base de dados bibliográfica.

Sistemas com arquivos bibliográficos e arquivos de autoridade integrados estão capacitados a prover um controle ativo sobre o desenvolvimento e a manutenção de uma base de dados. Por outro lado, sistemas com arquivos separados servem somente para orientar os catalogadores a criarem registros bibliográficos.

Dois importantes exemplos de sistemas de controle de autoridade operam na *Online Computer Library Center* (OCLC) e na *Washington Library Network* (WLN) duas das maiores redes bibliográficas existentes no mundo. A OCLC oferece busca e exibição de arquivos de autoridade de nomes e assuntos da *Library of Congress* (LC) através de seu sistema *on-line*. Esses arquivos não estão ligados a registros bibliográficos na *Online Union Catalog* da OCLC (OLUC). Entretanto, alguma consistência é obtida em cabeçalhos de nomes, incorporando registros MARC da LC (*Library of Congress*) no OLUC e pela ênfase das bibliotecas-membro em estabelecer cabeçalhos que sejam consistentes com as práticas da LC.[135]

A WLN, ao contrário da OCLC, mantém uma base de dados bibliográfica e de autoridade integrada e provê instalações para a verificação *on-line* de cabeçalhos em registros bibliográficos.[136]

---

[133] O'NEIL, E. T.; VIZINE-GOETZ, D. op. cit. p. 141.
[134] Burger *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 142.
[135] Taylor *et al apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 142.
[136] PEREIRA, Maria de Nazaré Freitas. *op. cit*.

Portanto, em última análise, o processo de validação entre registros bibliográficos e de autoridade encontrados nos sistemas de controle de autoridade contribuem de maneira significativa para a melhoria da qualidade dos arquivos bibliográficos nas bases de dados.

Detecção de duplicação

No contexto das bases de dados bibliográficas, registros duplicados são definidos como dois ou mais registros bibliográficos que representam o mesmo item bibliográfico.

A identificação de registros duplicados não é uma tarefa trivial. Esse tipo de erro ocorre devido a informações incorretas, incompletas ou ausentes, resultante de diferentes interpretações das regras de catalogação e de variações nas práticas de catalogação.[137]

Um grande número de registros duplicados pode afetar o desempenho da indexação e aumentar os custos e manutenção e armazenamento. Como será visto mais adiante, em detalhe, a duplicidade de registros pode levar, por exemplo, a resultados estatísticos incorretos como é o caso da base Currículo Lattes. Pela sua natureza cadastral ela armazena dados da produção científica de cada pesquisador cadastrado. Dessa forma, um artigo contendo quatro autores poderá produzir até quatro registros distintos para o mesmo item, supondo-se que os quatro autores são pesquisadores cadastrados na base. Evidentemente tal fato irá criar mais adiante distorções nos resultados de indicadores de produção científica construídos a partir desta importante base de dados.

A literatura aponta como causas prováveis da duplicidade de registros razões como a falta de cuidado nas pesquisas, dificuldades na edição e na atualização dos registros. Jones e Kastener[138] acreditam que uma das causas primárias da duplicidade de registros nas bases do OCLC e da RLIN (*Research Library Network*) seja a dificuldade dos catalogadores em distinguir reimpressões de edições quando consideram regras de catalogação, padrões de entrada de dados bibliográficos, mudanças nas tecnologias de impressão e práticas locais de catalogação.

---

[137] O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 144.
[138] Jones e Kastener *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 145.

Heller *et al*[139] discutem o problema de registros duplicados nas bases de dados do *Environmental Protection Agency/National Institutes of Health Chemical Information System* (CIS). Nesta base, os compostos químicos são identificados pelo nome, fórmula molecular e peso. A duplicidade de registros ocorre, nesse caso em particular, devido à variedade de nomes usados para o mesmo composto. Para eliminar a duplicidade de registros cada composto foi associado a um identificador único, denominado CAS *Register Number* – REGN.

Outra metodologia que merece destaque na detecção de duplicidade é a utilização de algoritmos de correspondência de registros. Essa abordagem apresenta diversas técnicas que identificam diferenças no conteúdo dos campos, elementos de dados ausentes e variações nas práticas de catalogação. Williams e Maclaury[140] desenvolveram um algoritmo em computador para identificação de registros duplicados. O algoritmo desenvolvido é um processo de duas etapas que, primeiro, reúne duplicatas em potencial usando uma chave título-data e, posteriormente, compara nomes, títulos e paginação. Testes com o algoritmo mostraram algumas falhas na identificação de duplicatas devido às variações encontradas nos títulos.

Hickey e Rypka[141] usaram uma chave de detecção duplicata em duas seções. Em uma primeira seção, denominada seção de correspondência exata, similar ao algoritmo descrito no parágrafo anterior, agrupam-se chaves relacionadas. A segunda seção consiste de campos-chave que poderiam ter uma correspondência exata ou parcial. Uma tabela de decisão é utilizada para determinar se as chaves são duplicatas. A aplicação do algoritmo detectou algo em torno de 60% dos registros duplicados. Apesar de não ser um percentual elevado, a simplicidade do algoritmo justifica sua implementação que pode ser realizada tanto retrospectivamente quanto numa verificação *on-line* de registros duplicados.

---

[139] Heller *et al apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 145.
[140] Williams e Maclaury *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 145.
[141] Hickey e Rypka *apud* O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 145.

Validação automática de dados e dados auto-verificáveis

A validação automática de dados é um conjunto de técnicas que consiste em detectar erros e corrigi-los automaticamente. Trata-se de uma técnica amplamente reconhecida e muito eficiente para garantir a qualidade das bases de dados. Os registros MARC são um bom exemplo do uso da validação automática de dados. Utilizando-se dos registros MARC vários tipos de erros podem ser detectados automaticamente através de: valores permitidos para designadores de conteúdo; padrões válidos de ocorrência para rótulos MARC; valores permitidos em determinados campos (p. ex., seqüências válidas de dados alfa-numéricos em campos de códigos de classificação). Em muitas situações, valores corretos ou *default* podem ser fornecidos automaticamente, baseados em outros dados contidos no registro ou em tabelas externas de valores.[142]

O conceito de dados auto-verificáveis (*self-checking data*) está baseado na adição de caracteres redundantes aos dados com o objetivo de facilitar a detecção de erros. Este conceito é também largamente utilizado nas tecnologias de informação, em particular, na área de telecomunicação digital que faz uso de dados auto-verificáveis nos protocolos de comunicação. Códigos de barra são um outro exemplo do uso de dados auto-verificáveis.

Os primeiros usos de dados auto-verificáveis em bases bibliográficas ocorreram na década de 60. O uso de caracteres de verificação tem sido usado com muito sucesso pelo *International Standard Book Number* (ISBN) e pelo *International Standard Serial Number* (ISSN). O ISBN é um número de dez dígitos usado para identificar de forma inequívoca uma publicação (monografia). Os primeiros nove dígitos identificam a publicação e o último dígito permite uma checagem automática sobre a validade do número ISBN, isto é, os primeiros nove dígitos. A verificação do número ISBN se dá através da soma dos produtos da multiplicação do primeiro dígito por 10, o segundo por 9 e assim sucessivamente até o nono dígito. Se o número ISBN for válido, o resultado da soma será exatamente divisível por 11. Se a divisão produz outro resultado, o ISBN é inválido.[143]

---

[142] O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 132.
[143] O'NEIL, E. T.; VIZINE-GOETZ, D. *op. cit*. p. 131.

3.4.2. O sistema da qualidade da OCLC

Com o objetivo de melhor ilustrar os conceitos e técnicas apresentados nas páginas anteriores, merece destaque uma descrição mais detalhada do sistema da qualidade que a OCLC adota e que se encontra minuciosamente descrito no capítulo 5 – "*Quality Assurance*" do guia "*Bibliographic Formats and Standards*" da OCLC.

A *Online Computer Library Center* - OCLC é a maior rede de computadores e telecomunicação de bibliotecas do mundo. É uma cooperativa sem fins lucrativos que oferece seus produtos e serviços a bibliotecas no mundo inteiro. A OCLC foi fundada em 1967 e inicialmente atendeu a 54 bibliotecas acadêmicas no estado de Ohio, EUA. Atualmente, mais de 43.000 bibliotecas de todos os tipos e tamanhos nos 76 países e territórios utilizam os produtos e serviços da OCLC. A OCLC é um dos mais antigos provedores de registros MARC. No ano de 2002 a base da OCLC contava com cerca de 48 milhões de registros.[144]

A garantia da qualidade dos serviços prestados pela OCLC é resultado da adoção de normas internacionais e da gestão de programas de controle de qualidade. Os programas de controle de qualidade da OCLC e de suas instituições-membro têm como principal objetivo a melhoria contínua dos registros da base WorldCat através da eliminação de registros duplicados e na correção de erros.

Para melhor compreensão do texto que se segue, faz-se necessário a descrição das definições de alguns produtos e conceitos utilizados pela OCLC:[145]

Bibliographic Formats and Standards Guide. Este guia refere-se exclusivamente a formatos e padrões estabelecidos para os registros eletrônicos de catalogação em *WorldCat* (nome comercial do OCLC *Online Union Catalog*). Ele estabelece convenções de rótulos (*tagging conventions*), padrões de entrada de dados e diretrizes para as informações que dão entrada no WorldCat.

---

[144] ABOUT OCLC. **Online Computer Library Center**. Disponível em: http://www.oclc.org/about/ . Acesso em: mar. 2003.

[145] OCLC. Introduction. In: **Bibliographic formats and standards guide**. Dublin, EUA: OCLC Online Computer Library Center, 2002. Disponível em: http://www.oclc.org/bibformats/en/introduction/. Acesso em: dez. 2002..

WorldCat (OCLC Online Union Catalog). Catálogo Coletivo Informatizado da OCLC. É uma base de dados de informações de catalogação e classificação. Seus registros são descrições bibliográficas eletrônicas de itens mantidos pelas instituições-membro da OCLC.

MARC (Machine-readable bibliographic records). Um registro bibliográfico eletrônico (MARC) consiste de campos. Um campo é uma área pré-definida na qual o mesmo tipo de informação bibliográfica é gravado. Os registros MARC na base *WorldCat* apresentam dois diferentes tipos de campos: campo fixo e campo variável.

Campo fixo. Um registro MARC possui um único campo fixo. Rótulos (*labels*) mnemônicos identificam os elementos que contém a informação codificada, descrevendo o item e o próprio registro.

Campo variável. Os demais campos em um registro MARC são variáveis no comprimento e no número. Cada campo variável pode ter de 1 a 1879 caracteres. Um campo variável MARC possui 3 partes:

- um rótulo de 3 dígitos.
- até dois indicadores de dígito único.
- um ou mais subcampos.

Rótulos (tags). Os rótulos MARC identificam os campos variáveis e são agrupados numericamente pela sua função.

Indicadores. Nos registros MARC, os indicadores dão informações sobre o campo para indexação, produção de fichas ou outras funções do sistema.

Subcampos. São as menores unidades lógicas de informação em um campo variável. Os códigos dos subcampos (letras ou números) identificam os subcampos e são precedidos por delimitadores de subcampo (‡). Subcampos normalmente contem a informação textual para a descrição bibliográfica do item, embora em alguns casos eles possam conter informação codificada.

Formatos bibliográficos. O sistema OCLC usa oito formatos MARC: Livros (BKS), Séries (SER), Material visual (VIS), Materiais mistos (MIX), Mapas (MAP), Escores (SCO), Gravações sonoras (REC), Arquivos de computador (COM).

A OCLC controla e corrige alguns dados de entrada em novos registros e os adiciona em registros existentes. O sistema OCLC inclui regras de validação de registros MARC para assegurar a entrada dos códigos e rótulos (*tags*) no padrão MARC. A OCLC também realiza uma varredura automática para corrigir dados obsoletos e incorretos na base WorldCat. Além disso, identifica e unifica registros em duplicata no formato BKS (livros) valendo-se de um *software* de detecção específico.

É importante ressaltar que as bibliotecas-membro da OCLC também são responsáveis pela precisão dos dados e pelo grau de adesão aos padrões de catalogação estabelecidos.

O capítulo "*Quality Assurance*"[146] do guia acima mencionado está dividido em 8 tópicos. Neles são apresentadas as abordagens e as estratégias que o sistema da qualidade da OCLC adota. São os seguintes:

1. Técnicas automatizadas.
2. Assistência do usuário.
3. Programas de cooperação.
4. Registros duplicados.
5. Relato de erros.
6. Submetendo relatórios.
7. Submetendo erros via o *Online System*.
8. Instruções e formulários.

Técnicas automatizadas

O OCLC adota basicamente duas técnicas automatizadas com o objetivo de garantir a integridade dos registros da base *WorldCat*. São elas: a varredura da base de dados e a utilização de um "software" de detecção e resolução de duplicidades.

[146] OCLC. Quality Assurance. In: **Bibliographic formats and standards guide**. Dublin, EUA: OCLC Online Computer Library Center, 2002. Disponível em: http://www.oclc.org/bibformats/en/quality/. Acesso em: dez. 2002.

A varredura da base *WorldCat* é feita com softwares que corrigem erros causados por mudanças nas regras de catalogação e nos padrões de entrada de dados.

O *software* de detecção e resolução de duplicidades identifica e unifica os registros duplicados no formato BKS. O *software* compara até 14 elementos descritivos bibliográficos dos pares de registros. Ele unifica os pares seletivamente baseando-se na similaridade dos elementos comparados. Vale ressaltar que o software só é aplicado apenas a um dos oito formatos MARC existentes na base. Mais adiante, será mostrado outras abordagens para registros duplicados.

Assistência do usuário

O usuário-membro do OCLC pode ele mesmo participar do sistema da qualidade de duas maneiras. A primeira, corrigindo erros e alterando seus próprios registros e, a segunda maneira, enriquecendo a base. No primeiro caso, o usuário só poderá fazer as modificações quando ele for o único "dono" daquele registro. No segundo caso, o usuário poderá enriquecer a base acrescentando dados em alguns campos "permitidos" desde que ele tenha um nível alto de autorização na catalogação da base.[147]

Programas de cooperação

As instituições-membro da OCLC participam de programas de cooperação com o objetivo de melhorar a qualidade da base de dados *WorldCat*. Neste sentido, destacam-se os programas ENHANCE, CONSER e o PCC (*Program for Cooperative Cataloging*).[148]

O programa ENHANCE foi criado para proporcionar a correção e adição de dados nos registros da base *WorldCat* para todos os formatos exceto o formato SER (séries).

O programa CONSER (*Cooperative Online Serials*) complementa o programa anterior. Ele proporciona a melhoria e a substituição dos registros referentes ao formato SER (séries).

---

[147] OCLC. Quality Assurance. *op. cit*. p. 2.
[148] OCLC. Quality Assurance. *op. cit*. p. 4.

Os objetivos do CONSER são dois: produzir e manter uma base de dados de registros de publicações seriadas de vários formatos e de amplo uso e contribuir para a catalogação e a criação de padrões dessas publicações. A base de dados do CONSER encontra-se inserida na base *WorldCat*.

O programa PCC (*Program for Cooperative Cataloging*) é um programa de cooperação internacional, coordenado pela LC (*Library of Congress*), juntamente com os participantes do PCC no mundo. É um projeto cujo principal objetivo é expandir o acesso a registros bibliográficos, proporcionando uma catalogação útil, rápida e de baixo custo orçamentário, seguindo regras e padrões comumente aceitos pelas bibliotecas em todo o mundo. O PCC é um programa que busca reduzir os custos dos participantes sem reduzir os padrões de catalogação e da qualidade como um todo.[149]

Registros duplicados

Registros duplicados são dois ou mais registros bibliográficos para o mesmo item. Na base da OCLC, registros duplicados são ocasionalmente permitidos, mas normalmente a duplicidade de registros é indesejável. No guia da OCLC a seção dedicada aos registros duplicados descreve como selecionar qual registro deve permanecer entre as demais duplicatas. A seleção é feita baseada em critérios pré-definidos e em função do tipo de formato bibliográfico que o registro faz referência.[150]

Relatando erros

Conforme visto anteriormente, em alguns casos, o usuário-membro da OCLC pode ele mesmo modificar, atualizar ou corrigir erros nos registros da base da OCLC. Em outros casos, entretanto, ele deverá relatar erros ou omissões à OCLC.

O guia da OCLC lista os tipos de erros que devem ser relatados e também os tipos de erros que não devem ser relatados. Notas e exemplos auxiliam o catalogador na

---

[149] PROGRAMA de catalogação cooperativa (PCC). Preparado pela equipe da Library of Congress Hispanic Reading Room. Disponível em: http://www.loc.gov/catdir/pcc/pccpor.html. Acesso em: nov. 2002.

[150] OCLC. Quality Assurance. *op. cit*. p. 5.

descrição dos tipos de erros listados. O guia também discrimina quais tipos de erros que exigem "provas" que devem acompanhar os relatos. A necessidade de "provas" ocorre para o relato de erros menos óbvios. Podem ser fotocópias de parte de um item bibliográfico ou documentos comprovando, por exemplo, a interrupção de um periódico ou a mudança de um título.

Fica a critério da OCLC o encaminhamento dos relatórios recebidos. A OCLC pode retornar o relatório à biblioteca para maiores esclarecimentos, ou pode encaminhá-lo para a biblioteca que criou o registro com o possível erro ou a outras bibliotecas para verificação. Relatórios ilegíveis, incorretos ou sem uma verificação adequada são descartados.

A forma de apresentação dos relatos de erros pode ser realizada de diversas maneiras. A OCLC oferece várias opções de envio que dependem da preferência do remetente e dos tipos de erros. Os relatórios de erros podem ser enviados através de carta, de formulários próprios enviados por fax ou e-mail, de formulários que podem ser encontrados no *site* da OCLC e também pelo *Online System*. Este último método só é utilizado quando não se requer a apresentação da "prova".[151]

3.5. Qualidade do conteúdo das bases de dados

Os novos paradigmas decorrentes da evolução tecnológica da informática e das telecomunicações nas últimas duas décadas propiciaram o surgimento de outras dimensões possíveis na percepção do usuário sobre a qualidade das bases de dados. Uma dessas dimensões é o "conteúdo" informacional. É a informação propriamente dita encapsulada na base de dados.

Para Sayão[152], até há poucos anos atrás, o conceito de qualidade de dados situava-se num plano essencialmente "físico". Isto é, a problemática estava centrada nas questões relacionadas à detecção de erros e na automação de procedimentos e técnicas de identificação e eliminação dos erros.

---

[151] OCLC. Quality Assurance. *op. cit*. p. 6.
[152] SAYÃO, L. F. Bases de dados e suas qualidades. In: LUBISCO, N.; BRANDÃO, L. (Ed.). **Informação e Informática**. Salvador: EDUFBA, 2000.

Pereira[153] destaca as questões relativas à qualidade do conteúdo, ou seja, as etapas e os procedimentos envolvidos na produção dos conteúdos veiculados pelas bases e que antecedem ao funcionamento destas.

A noção de qualidade para a informação propriamente dita está relacionada a conceitos tais como exatidão, atualização, novidade e consistência. Estes conceitos estão ligados a fatores que antecedem ao funcionamento das bases de dados, como por exemplo:

- Confiança nas fontes geradoras da informação – instituições produtoras, autores, pesquisadores, bibliotecários, editoras;

- Estratégia de coleta da informação – o escopo e a abrangência da coleta são medidas de qualidade. Quando, por exemplo, um produtor de bases de dados se propõe a cobrir toda a literatura produzida sobre um determinado assunto numa determinada língua, região ou país, isto se torna um compromisso relacionado à qualidade da base de dados.

- Seleção – metodologias para avaliação dos dados a serem incorporados.

Além da informação propriamente dita, Sayão[154] relaciona ainda dois outros aspectos ao "conteúdo informacional": a estrutura e a representação.

A estrutura da informação é definida no projeto da base de dados. Deve refletir o recorte de uma realidade e as necessidades de um universo de usuários reais ou postulados.

Os esquemas de representação da informação têm um impacto importante sobre a qualidade percebida pelo usuário, pois, segundo Sayão, influenciam diretamente na recuperação da informação. A representação depende da excelência dos indexadores catalogadores e dos instrumentos como os tesauros, listas de autoridades e esquemas de classificação. Manuais e normas são instrumentos especialmente importantes para as bases bibliográficas e catalográficas que podem operar em cooperação.

Pereira, tomando por base um trabalho de Rittberger e Rittberger, apresenta um conjunto de requisitos para a produção de conteúdos de qualidade. Segundo os autores,

---

[153] PEREIRA, Maria de Nazaré Freitas. *op. cit*. p.17
[154] SAYÃO, L. F. *op. cit.*

o conteúdo de uma base de dados tem que se orientar por registros de qualidade e testes aplicados às etapas de seu processo de produção. Os requisitos incluem: escopo e cobertura da área de assunto, abrangência, atualidade, precisão e consistência. Estes requisitos operam sobre as várias etapas que vão desde a aquisição do documento ao sistema de registro e de produção.[155]

A seguir, os requisitos propostos por Rittberger e Rittberger são descritos por Pereira do ponto de vista de uma base de dados bibliográfica:

1. *Escopo e cobertura da área de assunto. Estão diretamente relacionados à coleção de informações da base. Uma coleção pode cobrir o conteúdo de um assunto específico, de uma missão ou ser multidisciplinar. A abrangência geográfica, o idioma e a época de publicação são também considerados critérios de cobertura.*

2. *Abrangência. Uma coleção pode ser representada por todos os tipos de publicações e/ou documentos: monografias, dissertações, capítulos e artigos de monografias, periódicos científicos, artigos de periódicos, relatórios técnicos, anais e trabalhos apresentados em congressos, seminários e conferências, literatura cinzenta, patentes e normas. A coleção pode ser internacional, cobrir um ou mais países e ainda pode ser limitada por aspectos temporais ou lingüísticos.*

3. *Atualidade. Consiste no lapso de tempo decorrente entre a publicação de um texto (sua data de publicação) e a inserção desta publicação em uma base de dados.*

4. *Precisão. Significa evitar erros em todas as etapas de produção de uma base: na análise do documento, durante a entrada de dados nos campos, bem como erros ortográficos.*

5. *Consistência. Representa o grau de uniformidade praticado no processamento de todas as unidades de informação. Para alcançar um alto nível de consistência, é preciso seguir regras e instruções de trabalho na seleção de*

[155] PEREIRA, Maria de Nazaré Freitas. *op. cit*.

documentos (varredura), na catalogação (regras de catalogação), assim como na classificação e indexação (esquema de classificação, tesauro, regras de indexação).[156]

3.5.1. Critérios de seleção de periódicos para a base de dados LILACS

Para ilustrar as questões relativas ao conteúdo, isto é, os requisitos de qualidade relativos às etapas que antecedem à construção da base, são apresentados a seguir os critérios de seleção que a base LILACS se utiliza para garantir a qualidade do conteúdo que ela disponibiliza a seus usuários.[157]

A base LILACS – Literatura Latino-Americana e do Caribe em Ciências da Saúde, coordenada pela BIREME, compreende toda a literatura relacionada às Ciências da Saúde, produzida por autores latino-americanos e do Caribe, publicada nos países da região da América Latina e Caribe, a partir de 1982.

Os critérios para seleção de periódicos foram definidos para a orientação dos editores e das unidades integrantes do sistema Latino-Americano e do Caribe de Informação em Ciências da Saúde.

São os seguintes os critérios que a base LILACS considera no seu processo de seleção de periódicos:

1. *Conteúdo. O mérito científico de um periódico é o principal aspecto a ser considerado na seleção de um novo título. Para avaliação do mérito científico são considerados os seguintes fatores de qualidade: validade, importância, originalidade do tema, contribuição para a área temática em questão e a estrutura do trabalho científico.*

2. *Revisão por pares. A revisão e aprovação das contribuições para os periódicos devem ser realizadas pelos pares.*

---

[156] PEREIRA, Maria de Nazaré Freitas. *op. cit*. p.18

[157] BIREME. **Critérios de seleção de periódicos para a base LILACS**. São Paulo: BIREME, 2000. Disponível em: http://www.bireme.br/. Acesso em: nov. 2002.

3. *Comitê editorial. O periódico deve possuir um Comitê Editorial formado por especialistas com experiência reconhecida na área.*

4. *Regularidade de publicação. A regularidade é um critério obrigatório no processo de avaliação. O periódico deve ser publicado seguindo rigorosamente sua periodicidade pré-estabelecida.*

5. *Periodicidade. A periodicidade é um indicador do fluxo da produção científica da área específica coberta pelo periódico. Na área das Ciências da Saúde, segundo o critério LILACS é recomendado que o periódico seja, no mínimo, trimestral.*

6. *Tempo de existência. Para ser considerado no processo de avaliação do LILACS, o periódico já deve ter pelo menos quatro números publicados.*

7. *Normalização. O periódico deve ter especificado as normas de apresentação, estruturação dos textos e referências de modo que seja possível avaliar a obediência à normalização pré-estabelecida.*

8. *Apresentação gráfica. O periódico deve ter qualidade gráfica, isto é, padrões elevados de qualidade no que se refere ao projeto gráfico (layout), às ilustrações e à impressão*.[158]

3.6. Qualidade das bases de dados e a Internet

Apesar de não pertencer ao escopo do presente trabalho, a proximidade com o tema e a presença da Internet no mundo atual justifica uma breve abordagem. De certo modo e com algum cuidado pode-se encarar a Internet como uma gigantesca "base de dados" de bases de dados, uma parte dela organizada, e a outra, caótica.

Em 1995, o grupo SCOUG dedicou seu evento anual aos aspectos da qualidade na Internet. Neste encontro de profissionais da informação verificou-se a existência de algumas diferenças entre a indústria de bases de dados e a Internet, principalmente no tocante ao fato de que os provedores de informação na Internet não estavam muito

[158] BIREME. *op. cit*.

preocupados com aspectos financeiros. Isto significava que os provedores de informação tinham pouco ou nenhum incentivo para melhorar a qualidade de seus produtos. O SCOUG pôde também constatar que, por um lado, havia padrões técnicos bem estabelecidos como o HTML e outros padrões do *WWW Consortium* e do *Internet Engineering Task Force* (IETF) mas, por outro lado, não havia padronização do conteúdo. Levando em conta que o público da Internet ou parte dele já estaria disposto a pagar por um serviço com maior valor agregado, isto é, ter a seu dispor um serviço que daria acesso a "áreas catalogadas, seguras e de qualidade" na Internet, o grupo SCOUG identificou diversos aspectos de qualidade relacionados aos seguintes tópicos:[159]

1. Credibilidade.
2. Autoridade.
3. Indexação.
4. Registro.
5. Revisões / *ratings*.
6. Aspectos técnicos.
7. Segurança e privacidade.
8. "Feedback" / manutenção / assistência ao usuário.
9. Avisos de alerta.
10. Ajuda.
11. Direitos autorais e propriedade intelectual.
12. Ferramentas de busca.
13. *Download* confiável, transparente e padronizado.
14. Cobrança *on-line*.
15. Diretórios confiáveis de endereços de *sites*.
16. Aspectos de censura, auto-censura.
17. "Máquinas" de pesquisa.
18. Propaganda.
19. Mecanismos de pagamento.
20. Capacidade de rastrear uso do *site*.

[159] HOFMAN, P. *et al*. Specification for resource description methods Part 2: Selection criteria for quality controlled information gateways. In: **Project RE 1004 (RE): DESIRE – Development of a european service for information on research and education**. Deliverable D3.22, mar. 1996, 90p. Disponível em: http://www.ukoln.ac.uk/metadata/desire/quality/. Acesso em: nov. 2002. p.39.

21. Manutenção.

Observa-se, pela natureza do tema em pauta, que os aspectos de qualidade levantados pelo grupo SCOUG não se referem especificamente à qualidade da informação mas à Internet em geral.

Um relatório elaborado em 1995 pelo *Information Market Observatory* – IMO que trata de aspectos da qualidade de bases de dados comerciais identificou como sendo os principais problemas da Internet os seguintes:

1. Excesso de informação – freqüentemente redundante e imprecisa.
2. A ausência de um controle centralizado – inexistência da função editorial e de arbitragem.

Com relação à rede WWW, o relatório do IMO destaca o excesso de *sites* duplicados, a velocidade com que os *sites* aparecem e desaparecem, e o amplo espectro de tipos de informação que vão desde a mais relevante às triviais e obscenas. Pelo fato de não possuir qualquer tipo de controle de qualidade, torna-se difícil o trabalho dos pesquisadores na tarefa da recuperação da informação na Internet. O relatório conclui que a Internet *não se tornará uma ferramenta séria para os profissionais da informação até que os aspectos da qualidade estejam resolvidos*.[160]

Finalizando com as palavras de Ciolek:

> *Nossa maior loucura parece ser a nossa disposição em cultivar este sistema de comunicação global, aberto para tudo e para todos, sem primeiro assegurar que temos suficiente informação útil e confiável, precisa e atualizada para circular através de tal monstro (behemoth) em rede*.[161]

[160] HOFMAN, P. *op. cit*. p.40.
[161] Ciolek *apud* HOFMAN, P. *op. cit*. p.40.

4. MATERIAL E MÉTODO

4.1. Material

O sistema Currículo Lattes destaca-se entre os sistemas de bases de dados que compõem a Plataforma Lattes pela sua utilidade, abrangência e reconhecida aceitação pelos seus usuários. Ele é o formulário eletrônico do MCT, CNPq, FINEP e CAPES/MEC para o cadastro de dados curriculares de pesquisadores e de usuários em geral. O Anexo 1 apresenta um histórico do sistema Currículo Lattes.

A base Currículo Lattes (base CL) é a principal fonte de dados para construção de indicadores da produção científica e tecnológica do Brasil, disponibilizando, através dos mais de 320 mil currículos cadastrados na base (setembro de 2003), além de dados de identificação do pesquisador e de sua trajetória profissional, referências de artigos científicos publicados em periódicos e em anais de congressos, livros, patentes e diversos outros produtos provenientes das atividades de ciência e tecnologia desenvolvidas nas instituições de ensino e pesquisa do país.

Os dados contidos na base CL são utilizados para:

- avaliação da competência de candidatos à obtenção de bolsas e auxílios;
- seleção de consultores, de membros de comitês e de grupos assessores;
- subsídio à avaliação da pesquisa e da pós-graduação brasileiras.

A base CL é uma fonte de dados para aqueles que demandam informações para estudos e tomada de decisão em C&T, principalmente nos aspectos que envolvem o conhecimento coletivo sobre a produção intelectual dos cientistas que atuam no país e os seus mais variados desdobramentos.[162]

Apenas para citar alguns exemplos, os indicadores de C&T, construídos a partir da base CL, poderão influir na tomada de decisão sobre investimentos em pesquisa, no futuro de uma instituição ou na escolha de um cientista para um cargo estratégico.

[162] CNPq. Plataforma Lattes. Disponível em: http://lattes.cnpq.br/. Acesso em fev. 2003.

Fica evidenciado o importante papel que a base de dados Currículo Lattes representa para aqueles que se empenham na análise dos inúmeros aspectos que envolvem a gestão em C&T, sua relação com o desenvolvimento social e econômico e as repercussões em uma esfera política mais ampla. Trata-se, sem dúvida, nos dias atuais, da base de dados mais importante e mais abrangente que o sistema brasileiro de C&T possui no que se refere a dados sobre os pesquisadores brasileiros e suas respectivas produções científicas e técnicas. Portanto, essas são as razões que levaram à escolha da base Currículo Lattes como objeto de estudo da presente Dissertação.

4.2. Método

De modo atender ao tema do presente trabalho, isto é, a qualidade de bases de dados para a construção de indicadores de C&T, foi desenvolvida uma metodologia com o objetivo de avaliar o grau de precisão e confiabilidade dos dados contidos no Currículo Lattes e, por conseguinte, sua adequação como fonte primária de dados para a construção de indicadores de C&T precisos e confiáveis.

Propõem-se duas abordagens. A primeira busca avaliar a precisão e a consistência dos dados de entrada na base CL, isto é, os dados alimentados pelo pesquisador. A segunda abordagem visa avaliar a precisão dos indicadores gerados a partir dos dados da base CL. Tais indicadores são produzidos e disponibilizados através do subsistema da Plataforma Lattes denominado “Demografia Institucional”.

4.2.1. Avaliação dos dados de entrada na base Currículo Lattes

Pode-se afirmar que a qualidade dos dados de entrada, isto é, os dados que alimentam uma base, é o fator determinante que define a qualidade de uma base como um todo. Não por acaso que, ao se abordar esse assunto, a literatura sobre desenvolvimento de bases de dados sempre relembra um velho bordão: “se entra lixo, sairá lixo”. Isso quer dizer que, independentemente de todos os outros fatores e recursos necessários para a construção de uma base de dados, se a mesma for alimentada com dados imprecisos, desatualizados e inconsistentes, isto é, dados de baixa qualidade, inevitavelmente, os resultados que esta base irá fornecer serão também de qualidade inferior. Por isso, antes de mais nada, faz-se necessário avaliar a qualidade dos dados que dão entrada na base.

O presente estudo selecionou para avaliação os dados que compõem as referências bibliográficas das publicações produzidas pelo pesquisador e por ele registradas no seu Currículo Lattes. Tal escolha deve-se ao fato de que esses dados são utilizados na produção de um dos mais significativos conjuntos de indicadores de C&T, aqueles relacionados à produção científica.

Cada uma das referências bibliográficas selecionadas para o presente estudo foi copiada eletronicamente do currículo do pesquisador e "colada" em um arquivo de planilha eletrônica, permitindo, assim, a criação de uma única tabela contendo todas as referências utilizadas para o estudo e, nas colunas adjacentes, as observações pertinentes.

A avaliação de cada referência bibliográfica exige a definição de requisitos de qualidade que permitam, de algum modo, verificar o grau de qualidade da referência bibliográfica. A fidedignidade dos dados representa um dos requisitos de qualidade de uma base de dados e está associada à precisão e a confiabilidade dos dados contidos em um registro. Os indicadores usados para medir a fidedignidade das referências bibliográficas são os seguintes:

- completude
- correção
- normalização

A verificação da fidedignidade de cada referência bibliográfica só é possível confrontando-a com a fonte primária, isto é, comparando cada elemento informacional da referência bibliográfica com as informações contidas no artigo publicado. Portanto, foi necessário ter em mãos cada um dos artigos relacionados no conjunto de referências, objeto do presente estudo. A tarefa de recuperar os artigos relacionados na lista de referências foi entregue à equipe de bibliotecárias da instituição selecionada para o presente estudo. Vale ressaltar que, para não criar uma possível distorção nesse levantamento, os pesquisadores dessa instituição, autores dos artigos referenciados, não foram contatados. O propósito desse levantamento foi de se ter uma noção da capacidade de recuperação dos artigos referenciados a partir, exclusivamente, das informações contidas nesta lista de referências bibliográficas retiradas da base CL.

Apesar das condições favoráveis para a recuperação desses artigos, já que esta tarefa foi levada a cabo por bibliotecárias com grande experiência em lidar com a literatura especializada da área do conhecimento a qual a instituição de pesquisa se dedica, a tarefa não foi trivial na sua execução. Como destaca Pedrini[163], muitas instituições no Brasil dispõem de bases bibliográficas da sua produção científica, o que não significa que elas possuam os artigos referenciados os quais, em boa parte, estão dispersos nas inúmeras revistas estrangeiras especializadas (agravado pelo fato de que a manutenção das assinaturas dessas revistas vem sofrendo cortes ao longo dos últimos anos).

Para a análise das referências bibliográficas tomou-se como padrão a norma NBR-6023[164], versão de agosto de 2000, elaborada pela Associação Brasileira de Normas Técnicas (ABNT). Ela define "referência bibliográfica" como um conjunto padronizado de elementos descritivos, retirados de um documento, que permite sua identificação individual.

A norma NBR-6023 tem os seguintes objetivos:

1. Especificar os elementos a serem incluídos em referências.

2. Fixa a ordem dos elementos das referências e estabelece convenções para a transição e apresentação da informação originada do documento e/ou outras fontes de informação.

3. Orienta a preparação e a compilação de referências de material utilizado para a produção de documentos e para inclusão em bibliografias, resumos, resenhas, recensões e outros.

Outro importante documento utilizado como referência, elaborado pela Divisão de Documentação Técnica da CPRM (Companhia de Pesquisa de Recursos Minerais), foi o guia "Referências e citações bibliográficas: guia prático com exemplos em

---

[163] PEDRINI, A. G. **O cientista e os métodos de avaliação de seu desempenho: estudo de sua adequação no contexto brasileiro**. Orientador: Rosali Fernandez de Souza. Rio de Janeiro: UFRJ, Escola de Comunicação – CNPq/IBICT, 1999. 442p. Tese. (Doutorado em Ciência da Informação).

[164] ASSOCIAÇÃO BRASILEIRA DE NORMAS TECNICAS. **Informação e documentação – referências – elaboração: NBR 6023**. Rio de Janeiro, ago. 2000. 22p.

geociências"[165]. Trata-se de um documento, baseado na norma NBR 6023, que traz orientações para a normalização de referências bibliográficas pertinentes à área de Geociências descritas nas versões impressa e digital.

No caso particular do presente trabalho faz-se necessário definir precisamente os elementos descritivos de um tipo de documento: o artigo completo publicado em um periódico, pois este foi o tipo de documento que foi selecionado na base CL para o presente estudo.

Segundo a norma NBR 6023 os elementos que compõem uma referência bibliográfica de um artigo publicado em periódico são os seguintes:

Elementos essenciais:

- autor(es) (se houver),
- título do artigo ou matéria,
- subtítulo (se houver),
- título do periódico,
- local da publicação,
- numeração correspondente ao volume e/ou ano,
- fascículo ou número,
- paginação inicial e final do artigo ou matéria,
- as informações de período e data de publicação.

Exemplo:

> Sobral, L. G. S.; Granato, M. Palladium: extraction and refining. **Minerals Engineering**, Inglaterra, v. 5, n. 1, p. 17-25, 1992.

Uma vez definida uma especificação normalizada dos elementos descritivos essenciais que compõem uma referência bibliográfica de um artigo publicado em periódico, o passo seguinte foi definir o procedimento no qual cada referência seria submetida aos critérios de fidedignidade e classificadas segundo determinadas categorias.

---

[165] CPRM. **Referências e citações bibliográficas: guia prático com exemplos em geociências**. Rio de Janeiro: CPRM/DIDOTE, 2001. 28p.

Os passos adotados foram os seguintes:

1. Verificação se a referência remete a um documento do tipo “artigo completo publicado em periódico”.

2. Verificação se o título do periódico consta na base CCN, acessada através da página *web* do IBICT.

3. Verificação da completeza, da precisão e da normalização dos elementos descritivos da referência bibliográfica a partir do confronto com a norma NBR-6023 e com os dados extraídos da fonte primária, isto é, o próprio documento referenciado.

4. As observações pertinentes a cada referência são anotadas na coluna “observações” da tabela construída utilizando-se de um programa aplicativo de planilha eletrônica Microsoft Excel 2000.

5. Em função da avaliação realizada nos passos anteriores cada referência foi classificada segundo as categorias abaixo descritas. Para melhor visualização do resultado global da avaliação realizada adotou-se, na planilha eletrônica, uma cor distinta para cada categoria, aplicada no fundo de cada célula da planilha que continha o texto da referência bibliográfica.

São as seguintes as categorias aplicadas a cada referência analisada:

1. Referência completa e correta ou com pequena falta ou erro que não compromete sua recuperação.

2. Referência de uma série monográfica .

3. Referência de um artigo de anais de evento.

4. Referência incompleta ou com erros que comprometem sua recuperação. Na coluna do autor indica falta de 1 ou mais autores.

5. Referência incompleta ou com erros que não permitiram sua localização.

A categoria 1 foi atribuída à referência bibliográfica cujo documento referenciado era comprovadamente um artigo publicado em periódico, apresentava corretamente os elementos descritivos ou continha pequena falta ou erro que não comprometia seriamente a sua recuperação. Tais faltas ou erros podiam ser, por exemplo, a falta da paginação completa, a não indicação do local da publicação e pequenos erros de ortografia. A falta de pelo menos um dos demais elementos descritivos excluía a referência dessa categoria.

As categorias 2 e 3 foram criadas em função de dois equívocos cometidos com uma certa freqüência pelo alimentador da base. Trata-se da classificação equivocada, feita pelo pesquisador, de referências relacionadas a séries monográficas e artigos publicados em anais de eventos, ou seja, referências que foram incluídas na seção do Currículo Lattes dedicada somente a artigos completos publicados em periódicos e que, muitas vezes, apesar de estarem completas e precisas, não correspondem ao tipo de documentado a ser incluído na referida seção.

Na categoria 4 foram identificadas aquelas referências cujos elementos descritivos se apresentavam incompletos e/ou com erros que comprometiam seriamente a recuperação do documento referenciado, incluindo, também, nessa categoria quando da falta de um ou mais autores do documento referenciado.

A categoria 5 reuniu todas as referências que apresentavam as mesmas características encontradas na categoria 4 e, acrescido do fato de que, durante o período de tempo em que as bibliotecárias se dedicaram à recuperação dos documentos, estes não foram efetivamente recuperados.

4.2.2. Avaliação dos indicadores gerados pelo sistema Demografia Institucional

Dentre os indicadores de produção científica, aqueles relacionados à produção bibliográfica destacam-se entre os mais importantes e os mais utilizados pelos estudiosos das áreas de política e gestão em C&T na avaliação e na comparação de pesquisadores, áreas do conhecimento, departamentos, instituições e regiões geográficas.

O sistema Demografia Institucional é um sistema aplicativo da Plataforma Lattes que apresenta uma série de indicadores referentes à pesquisa e produção científica, tecnológica e artístico-cultural, segundo departamentos, centros ou áreas de atuação dos autores. Seu principal objetivo é apresentar o perfil da pesquisa e produtividade de professores, pesquisadores, alunos e demais pessoas vinculadas a uma instituição, a partir de critérios configuráveis pelo usuário do sistema, quanto à distribuição da população pesquisada.[166] Os dados primários utilizados para a construção destes indicadores são extraídos da base CL. Vale destacar que, apesar do Sistema Demografia Institucional estar disponibilizado ao público, via Internet, encontra-se na sua página de consulta um aviso que informa aos usuários do mesmo que o sistema "encontra-se em fase de testes e validações finais - versão Beta" (http://lattes.cnpq.br - consulta em fev. 2003).

O acesso ao sistema Demografia Institucional se faz, inicialmente, selecionando a instituição de interesse. Em seguida, o sistema oferece ao usuário telas de menus nas quais são definidas as unidades de análise, as variáveis de corte e os filtros desejados. Para o presente estudo foi solicitado ao sistema as informações e os resultados dos indicadores relativos à produção bibliográfica anual da instituição selecionada. Como resultado da pesquisa o sistema Demografia Institucional gerou dois tipos de relatório. O primeiro, apresentava no formato de uma tabela, os seguintes resultados:

- numeração seqüencial correspondente a cada referência bibliográfica recuperada pela pesquisa;
- o nome do autor pelo qual o sistema associa a referência à instituição pesquisada;
- título do artigo publicado;
- ano de publicação do artigo;
- país onde é editado o periódico que publicou o artigo.

O segundo relatório apresentava as referências bibliográficas na sua forma completa, com a descrição de todos os elementos essenciais e outras informações pertinentes ao

[166] CNPq. Plataforma Lattes. Disponível em: http://lattes.cnpq.br/. Acesso em fev. 2003.

assunto do artigo, como palavras-chave, grande área e subárea do conhecimento, setores de atividade e meio de divulgação.

Para a análise das referências foi aplicada uma metodologia que consistiu em submeter cada referência bibliográfica gerada pela consulta ao sistema Demografia Institucional aos seguintes passos:

- identificação dos autores do documento com o propósito de determinar se o autor era pesquisador atuante na instituição selecionada no ano da pesquisa. Todos os autores os quais o sistema associava a referência à instituição selecionada tiveram seus Currículos Lattes consultados.

- identificação do tipo de documento referenciado para efeito da contagem de apenas artigos publicados em periódicos.

- identificação das referências duplicadas de artigos publicados em periódicos.

Portanto, uma vez aplicada a presente metodologia sobre a lista de referências bibliográficas, os novos resultados obtidos para os indicadores estudados foram devidamente tabulados de modo permitir o confronto destes resultados revisados com os resultados originais fornecidos pelo sistema Demografia Institucional e com os resultados oficiais fornecidos pela instituição selecionada.

4.3. Amostra

Por razões de praticidade e da necessidade de limitar o escopo da pesquisa fez-se necessário definir um corte na base CL de modo a restringir, para efeito do presente estudo, o conjunto de dados a serem avaliados. O primeiro critério para a seleção da amostra a ser estudada foi a escolha de uma instituição de pesquisa científica que oferecesse amplas facilidades para o acesso às informações necessárias ao presente estudo. Além disso, seria muito importante para o estudo proposto conhecer o contexto do ambiente institucional nos seus variados aspectos como, por exemplo, no tocante ao pesquisador e as relações da sua produção bibliográfica com os meios de divulgação, os títulos dos periódicos e os eventos mais expressivos da área, a terminologia específica, etc. Portanto, levando-se em consideração esses fatores, a escolha recaiu sobre o Centro

de Tecnologia Mineral – CETEM, instituição na qual o autor da presente Dissertação atua há quase duas décadas, inicialmente, como pesquisador e, nos últimos 7 anos, como responsável pela área de informação a qual, no CETEM, compreende os setores de informática, editoração, biblioteca e divulgação técnica. O Centro de Tecnologia Mineral – CETEM, localizado no Rio de Janeiro, é um centro de pesquisas, subordinado ao Ministério da Ciência e Tecnologia – MCT, dedicado à pesquisa científica e tecnológica nas áreas minero-metalúrgica, meio ambiente e economia mineral. No Anexo 2 encontra-se um breve histórico da referida instituição.

O segundo critério adotado para a definição da amostra foi a escolha do tipo de produto da atividade científica a ser estudado. Pela sua importância e, a princípio, disponibilidade, foi selecionado o artigo científico publicado em periódicos e em anais de eventos científicos. A seguir, são apresentados os demais critérios que permitiram definir as amostras utilizadas nas distintas abordagens propostas.

4.3.1. Dados de entrada na base Currículo Lattes

Os dados de entrada da base CL foram extraídos dos currículos dos pesquisadores atuantes no Centro de Tecnologia Mineral – CETEM no ano de 2000, totalizando 54 currículos. A obtenção desses currículos foi realizada acessando-se, via Internet, a página *Web* do CNPq (www.cnpq.br). À medida que cada currículo era totalmente disponibilizado no microcomputador local procedia-se à gravação do mesmo no formato HTML, possibilitando, assim, "congelar" numa determinada data um conjunto de currículos os quais, posteriormente, seriam objeto deste estudo. A título de ilustração, o Anexo 3 apresenta a estrutura de apresentação dos dados do Currículo Lattes de um determinado pesquisador. A data em que se realizou esta operação de levantamento de currículos foi a de 31 de março de 2001.

Desses 54 currículos, foram selecionadas para estudo as referências bibliográficas registradas pelos pesquisadores na seção do Currículo Lattes destinada aos "artigos completos publicados em periódicos". No Anexo 4 deste trabalho encontra-se a tabela impressa, construída originalmente em planilha eletrônica Excel 2000, apresentando as 235 referências bibliográficas, objeto do presente estudo.

4.3.2. Indicadores gerados pelo sistema Demografia Institucional

Com o objetivo de avaliar a qualidade dos resultados dos indicadores de produção científica que a Plataforma Lattes disponibiliza através do sistema Demografia Institucional, o presente trabalho selecionou para esse estudo de caso um conjunto de indicadores da produção bibliográfica do CETEM, aplicando-se os seguintes critérios:

- Produção de C&T da instituição: ”CETEM”;
- Tipo de produção bibliográfica: “artigos publicados (em periódicos) no país”; “trabalhos publicados em eventos”
- Ano da produção: “2000”; “2001”; “2002”.

Dessa forma, aplicados os critérios acima descritos, foram definidos dois conjuntos de indicadores, o primeiro correspondendo ao número de artigos em periódicos publicados pelos pesquisadores do CETEM nos anos de 2000, 2001 e 2002 e o segundo correspondendo ao número de trabalhos publicados em eventos pelos pesquisadores do CETEM nos anos de 2000, 2001 e 2002. Portanto, conforme apresentado no Quadro 2 abaixo, a presente amostra compõem-se de 740 registros de referências bibliográficas.

Quadro 2. Produção científica do CETEM

Ano	Artigos em periódicos	Trabalhos em eventos	Total
2000	77	236	313
2001	85	251	336
2002	55	36	91
			740

Fonte: Plataforma Lattes – sistema Demografia Institucional. Anexo 6

5. RESULTADOS

A seguir, são apresentados os resultados da avaliação da qualidade dos dados da base CL submetidos à metodologia descrita no capítulo anterior.

5.1. Dados de entrada na base Currículo Lattes

A amostra avaliada era composta de 235 referências bibliográficas. Cada uma dessas referências foi submetida à metodologia descrita no item 4.2.1. No Quadro 3 abaixo encontra-se o resultado quantitativo da referida análise, em números absolutos (N) e em percentuais (%).

Baseado na metodologia proposta, a situação ideal seria algo próximo do percentual de 100% das referências estudadas classificadas na categoria 1, ou seja, todas as referências apresentando seus principais elementos descritivos de forma completa, precisa e atendendo às especificações da norma NBR-6023. Como já anteriormente mencionado, pequenas faltas ou incorreções foram admitidas nas referências classificadas na categoria 1 pelo fato de não comprometerem seriamente a recuperação da publicação. Essas pequenas incorreções admitidas foram as seguintes: a falta do local de publicação, a falta da paginação e pequenos erros de ortografia.

A falta ou a imprecisão dos dados referentes aos demais elementos descritores, a saber: nome dos autores, títulos do artigo e do periódico, volume, número e ano de publicação, foram considerados, nesta metodologia, como passíveis de provocar um maior comprometimento na recuperação da publicação referenciada. Portanto, as referências com estas características foram classificadas nas categorias 4 e 5.

O grande número de referências equivocadamente classificadas pelo pesquisador como artigo de periódico, em particular, artigos de anais de evento e séries monográficas, fez merecer destaque e, por conseguinte, a criação das categorias 2 e 3.

Quadro 3. Resultado da classificação das referências bibliográficas analisadas

Categoria	Descrição	N	%
1	Referência completa e correta ou com pequena falta ou erro que não compromete sua recuperação.	84	35,7
2	Referência de uma série monográfica.	43	18,3
3	Referência de um artigo de anais de evento.	24	10,2
4	Referência incompleta ou com erros que comprometem sua recuperação (inclui a falta de 1 ou mais autores).	55	23,5
5	Referência incompleta ou com erros que não permitiram sua localização.	29	12,3
	Total	235	100

Fonte: Anexo 4

Dessa forma, nesta pesquisa, 35,7% das referências analisadas encontravam-se na categoria 1, ou, de outra forma, 64,3% das referências não se enquadraram nos padrões de qualidade desejáveis, especificados na categoria 1.

As referências que não remetiam a artigo publicado em periódico representaram 28,5% do total, distribuídas nas categorias 2 (18,3%) e 3 (10,2%).

As categorias 4 e 5 reúnem as referências que apresentaram erros ou faltas que comprometiam a recuperação do artigo referenciado. A diferença entre estas duas categorias reside no fato de que as referências classificadas na categoria 4 puderam ser localizadas e as da categoria 5 não foram localizadas. Esse subconjunto totalizou 35,8% do universo de referências estudadas ou, ainda, 23,5% na categoria 4 e 12,3% na categoria 5, respectivamente.

O Quadro 4 apresenta uma outra forma de agrupar os resultados obtidos. Este reagrupamento permite classificar as referências analisadas em 3 critérios distintos. A categoria 1 agrupa as referências que atendem as especificações mínimas de qualidade, adotando-se a metodologia proposta. O agrupamento das categorias 4 e 5 representa o

conjunto de referências que não atendem às especificações mínimas e o agrupamento das categorias 2 e 3 representa o conjunto de referências que não remetem a artigos de periódicos.

Quadro 4. Critérios de qualidade das referências bibliográficas

Categoria	Critério	%
1	Atendem às especificações mínimas	35,7
4 + 5	Não atendem às especificações mínimas	35,8
2 + 3	Não remetem ao tipo de documento especificado	28,5
	Total	100,0

Fonte: Anexo 4

Excluindo-se as referências que não remetem a artigos de periódicos (categorias 2 e 3), observa-se que o restante das referências ocorre em percentuais praticamente iguais entre as que atendem (categoria 1) e as que não atendem (categorias 4 e 5) às especificações mínimas, nesta avaliação em particular.

5.2. Indicadores gerados pelo sistema Demografia Institucional

Para quem conhece o contexto da atividade de pesquisa da instituição selecionada para o presente estudo, no caso, o CETEM, percebe-se, de imediato, ao observar os resultados, fornecidos pelo sistema Demografia Institucional para os indicadores de produção bibliográfica, que os valores estão fora da realidade da instituição. Os valores apresentados mostram-se elevados e indicam que as causas para esse problema parecem ir além dos equívocos causados pelo pesquisador na entrada de dados na base CL ou devido à duplicação de referências. A análise cuidadosa de cada referência mostrou que um grande número delas, na verdade, não pertencia à produção científica da instituição selecionada. Verificou-se uma característica comum entre essas referências erradamente identificadas pelo sistema como produção do CETEM: pelo menos um dos autores

encontrados nestas referências foi, durante um certo período de tempo, pesquisador do CETEM. Tal constatação exigiu um trabalho redobrado para a correta identificação destas referências. Não bastava conferir se o pesquisador era do CETEM mas, também, foi necessário verificar em que período ele atuou como pesquisador no CETEM.

Inicialmente foram identificados, em cada referência, os autores que atuaram no CETEM nos últimos três anos (2000, 2001 e 2002), período da pesquisa. Em seguida foi identificado o autor que o sistema Demografia Institucional associava a referência ao CETEM. O Currículo Lattes de cada um desses autores foi consultado e foi constatado que a grande maioria, com exceção de dois deles, em algum momento da sua vida profissional atuou no CETEM e essa sua passagem estava registrada no seu Currículo Lattes. Portanto, os resultados gerados pelo sistema Demografia Institucional mostrou uma grave falha na sua lógica de busca, pois, ao solicitar ao sistema a produção bibliográfica de uma instituição, num determinado ano, o sistema apresentava não apenas a produção bibliográfica dos pesquisadores atuantes na instituição naquele ano, mas também, acrescentava toda a produção bibliográfica daquele ano de um grande número de pesquisadores que já teriam atuado no CETEM em algum período de tempo e, portanto, não atuavam no CETEM no ano da pesquisa.

Tomando-se de um exemplo para melhor entendimento, a referência abaixo é apontada pelo sistema Demografia Institucional como produção bibliográfica do CETEM no ano 2000:

> ALBAGLI, Sarita. Amazonie: frontière géopolitique. **Corrier de la Planète**, Paris, v. 6, n. 60, 2000.

A pesquisadora Sarita Albagli atuou no CETEM no período de 1989-1994. Posteriormente ela veio a trabalhar no IBICT, atuando neste instituto até os dias atuais. Como se pode observar, um produto que ela gerou no IBICT em 2000 está sendo também contabilizado como produção do CETEM no mesmo ano.

Assim como ocorre com a pesquisadora acima referida, outros 25 pesquisadores (ano 2000), que já atuaram no CETEM, aparecem, também, com suas respectivas produções bibliográficas de suas respectivas instituições onde atuam, como produção do CETEM,

provocando um acréscimo totalmente equivocado no resultado do indicador ora em estudo. Portanto, além do sistema contabilizar referências duplicadas em função do número de autores de um artigo de periódico, ocorre também uma duplicação de referências em função do número de instituições onde o pesquisador atuou.

Os resultados obtidos podem ser visualizados no Quadro 5. A segunda coluna do quadro apresenta o resultado numérico fornecido pelo sistema Lattes. A terceira coluna, denominada "Lattes Revisado", apresenta o resultado após a aplicação da metodologia acima descrita. Com o objetivo de conferir maior credibilidade ao estudo, a quarta coluna do Quadro 5, denominada "CETEM", apresenta os mesmos indicadores cujos resultados foram extraídos dos documentos oficiais da instituição.

Quadro 5. Número de artigos publicados em periódicos pelo CETEM

Ano	LATTES[1]	LATTES REVISADO[2]	CETEM[3]	Erro* %
2000	77	19	20	285,0
2001	85	19	21	304,8
2002	55	16	21	161,9

Fonte:
1. Plataforma Lattes - Sistema Demografia Institucional. Anexo 5
2. Anexo 6.
3. Relatórios de gestão do CETEM.
* percentual de erro em relação ao valor CETEM.

Observa-se de imediato que a proximidade dos valores apresentados na terceira e quarta colunas da tabela indicam que esses valores devem ser os mais próximos da realidade para os resultados dos indicadores estudados. Uma maior diferença encontrada na comparação entre os valores apresentados para ano de 2002 pode ser explicada pelo fato de que, muito provavelmente, alguns pesquisadores ainda não teriam atualizado seus CLs à época da pesquisa, janeiro de 2003. Por outro lado, a base de dados do CETEM que registra a sua produção científica, por exigências internas, já se encontrava atualizada para efeito de geração de relatórios de gestão de encerramento do ano e da avaliação de desempenho dos pesquisadores.

Sem aplicar a metodologia utilizada nos artigos publicados em periódicos mas constatando-se através de uma análise superficial que os mesmos problemas descritos nos parágrafos anteriores se repetem também para os resultados relativos à produção de trabalhos publicados em anais de eventos, foi possível fazer uma comparação dos valores para este indicador gerados pelo sistema Lattes com os valores oficiais apresentados nos relatórios de gestão elaborados pelo CETEM. O Quadro 6 abaixo mostra os resultados comparativos coletados nas duas fontes referidas.

Quadro 6. Trabalhos publicados em anais de eventos pelo CETEM (inclui resumos)

Ano	LATTES[1]	CETEM[2]	Erro %
2000	236	57	314,0
2001	251	84	198,8
2002	36	79	-54,4

Fonte:
1. Plataforma Lattes - Sistema Demografia Institucional.
2. Relatórios de gestão do CETEM.
* percentual de erro em relação ao valor CETEM.

Comparando-se os valores apresentados no quadro acima, observa-se novamente uma grande disparidade entre os mesmos. Chama a atenção os valores relativos ao ano de 2002. A princípio parece inverter a tendência observada nos outros anos. Conhecendo o contexto das atividades científicas do CETEM é possível dar uma explicação razoável para esse detalhe observado. Um grande número de referências não tinha, ainda, dado entrada na base CL pelo fato de que um grande evento da área de Tecnologia Mineral ocorrera em novembro de 2002 e a pesquisa realizada no sistema Lattes tenha sido realizada em fevereiro de 2003. Por outro lado, o valor apresentado pelo CETEM está muito próximo da realidade pelo fato de que o pesquisador do CETEM tinha uma motivação muito forte para atualizar a base de dados do CETEM com a sua produção científica. Essa motivação deve-se ao fato de que a avaliação de desempenho do pesquisador afetaria diretamente no cálculo da GDACT, uma gratificação que incide

sobre o salário do pesquisador em função da sua produtividade. Essa avaliação, realizada em dezembro de 2002, exige que o pesquisador tenha sua produção científica mais recente atualizada no banco de dados da instituição. Portanto, naquele momento específico de final de ano, o pesquisador apressou-se em atualizar a base institucional e deixou para mais tarde a atualização do seu CL. Vale lembrar que o autor desta Dissertação é o responsável pela gestão dos sistemas de informação do CETEM e, portanto, acompanha de perto as demandas das coordenações e da diretoria do Centro por informações gerenciais.

Esse relato mostra, de certa forma, que o conhecimento do contexto de onde os dados são gerados pode contribuir para a qualidade dos dados, o que reforça a tese de que a necessidade de validação dos dados, ou pelo menos parte dessa validação, deveria ser feita de forma descentralizada pelas instituições.

6. CONSIDERAÇÕES FINAIS

A análise dos resultados obtidos na avaliação dos dados de entrada permite identificar ou sugerir algumas causas que poderiam explicar, pelo menos em parte, o percentual elevado de referências fora das especificações mínimas de qualidade, estabelecidas pela metodologia proposta neste estudo. Inicialmente, faz-se necessário ressaltar alguns aspectos do contexto que envolve a base Currículo Lattes. Em primeiro lugar, sabe-se que, de uma maneira geral, o pesquisador mostra uma certa relutância em preencher formulários, incluindo aí o seu próprio CL. Além disso, o não preenchimento não trazia maiores conseqüências para o pesquisador, pelo menos até o ano de 2000. É bom lembrar que os currículos analisados nesta pesquisa foram coletados da base CL em março de 2001. Naquela data a base CL tinha sido disponibilizada ao público em geral havia pouco mais de dois ou três meses. É provável que, naquela data, poucos pesquisadores conheciam este fato. Portanto, a percepção do pesquisador sobre a importância da base para o sistema de C&T era relativamente baixa e tal fato poderia contribuir para uma certa displicência no preenchimento do seu CL, afetando a qualidade dos dados que são alimentados à base.

É de se supor que, a partir da disponibilização da base CL ao público e de outros acontecimentos relevantes, como a vinculação da base CL ao Diretório dos Grupos de Pesquisa, a recente criação do sistema Demografia Institucional e a necessidade do pesquisador obrigatoriamente manter atualizado seu currículo eletrônico para que ele possa concorrer a bolsas e participar de projetos de pesquisa fomentados pelo governo, o pesquisador venha mudando, nos últimos anos, sua percepção com relação à importância da base CL. Tal mudança de percepção poderia estar refletindo em um maior cuidado por parte do pesquisador no momento do preenchimento do seu CL o que poderia significar uma melhoria na qualidade dos dados que alimentam a base CL. A constatação dessa mudança de percepção poderia ser tema de um estudo a ser realizado no futuro. Tal estudo poderia comparar currículos eletrônicos obtidos em anos distintos e submetê-los à metodologia utilizada nesta Dissertação. Os valores dos percentuais obtidos para as categorias em cada ano poderiam fornecer indicações que demonstrassem a variação da percepção do pesquisador em relação à importância e às finalidades da base CL.

Retornando aos resultados da análise das referências, observa-se uma percentagem elevada (28,5%) de referências de trabalhos publicados em anais de evento e séries monográficas classificadas como artigo publicado em periódico. Entre as causas prováveis que explicam esse fato, certamente, uma certa displicência ou desinteresse do pesquisador no preenchimento do seu CL, como já mencionado, pode ser uma delas. No caso do trabalho publicado em anais de evento, é difícil justificar tal equívoco, já que na grande maioria destas referências está claramente explícito o nome do evento. Por outro lado, a série monográfica pode induzir o pesquisador ao erro, tanto pelo fato de poder ser confundida pelo leigo com a série periódica, isto é, a revista ou periódico e também pelo fato do CL não prever esse tipo de classificação de publicação. Tecnicamente, as séries monográficas devem ser classificadas juntamente com os livros por se tratarem de monografias. Tal afirmação pode ser confirmada ao se consultar o guia "Referências e citações bibliográficas: guia prático com exemplos em geociências", elaborado pela CPRM e que cita como exemplo de monografia um título de uma das séries do CETEM.

A classificação equivocada do tipo de publicação não chega a afetar de maneira significativa a recuperação de um determinado documento. No entanto, quando a finalidade da base também é a de ser fonte primária para construção de indicadores de produção científica, esta finalidade fica bastante prejudicada, principalmente, no que se refere a contabilização dos diversos tipos de produtos bibliográficos gerados pela atividade científica. A comunidade científica, e a estrutura do CL demonstra isso, considera importante a diferenciação dos tipos de publicação que o pesquisador produz, seja um artigo de periódico, seja um capítulo de livro ou um relatório de projeto, apenas para citar alguns exemplos. Além disso, e de uma forma bem subjetiva, a comunidade atribui pesos distintos quanto à importância de um produto bibliográfico em relação a outro. Para citar um exemplo mais concreto, verifica-se que, em geral, é atribuído um valor maior ao artigo publicado em periódico (e mais ainda se o periódico for indexado) do que ao trabalho publicado em anais de evento. Essa importância relativa varia em função da área científica que se está investigando. Por exemplo, na Física, os pesquisadores atribuem um valor maior ao artigo publicado em periódico. Já na área de tecnologia mineral essa distinção é menos evidente, até porque nessa área publica-se mais em anais de eventos.

Essa constatação, a importância maior que a comunidade científica confere ao artigo publicado em periódico, pode também explicar, pelo menos em parte, o percentual elevado de referências que não são artigos publicados em periódico e que são classificados como tal (28,5% to total desta pesquisa). É possível que um ou outro pesquisador possa considerar sua produção de artigos em periódicos insuficiente e, ao preencher o seu CL, ele desloque algumas referências com o objetivo de aumentar o valor numérico de artigos publicados em periódicos no seu CL. Vale lembrar que, conforme já comentado anteriormente, o CL vem sendo utilizado, cada vez mais, na avaliação de pesquisadores que se candidatam a bolsas de produtividade e outros benefícios.

Feito essas considerações, torna-se possível sintetizar algumas das possíveis causas que afetam negativamente a qualidade dos dados que são alimentados na base CL, a saber:

- Baixa percepção do pesquisador quanto à importância e as finalidades da base CL.

- Refratariedade cultural da comunidade científica a preencher formulários (burocracia desnecessária e exercício de um tipo de "controle").

- Desconhecimento das regras básicas de normalização de referências bibliográficas e sua importância para a recuperação do documento e para a construção de indicadores de produção científica.

- A interface de interação usuário/sistema Lattes ainda é relativamente complicada e confusa.

Uma vez detectadas as possíveis causas que contribuem para a baixa qualidade dos dados que alimentam a base CL, é possível sugerir algumas ações visando melhorar a qualidade dos mesmos.

- Incrementar junto aos pesquisadores a divulgação da importância e das finalidades da base CL.

- Implementar procedimentos de avaliação da base.

- Explicitar especificações técnicas e de qualidade (rótulos).

- A partir das estatísticas de erros encontrados na alimentação da base, disponibilizar informações ao pesquisador com orientações visando minimizar os erros mais freqüentes encontrados.

- Criar uma instância que valide os dados de entrada na base. Essa instância poderia ser a instituição onde atua o pesquisador, seja através de um serviço de informação ou através da biblioteca da instituição, setores mais preparados para assumir essa tarefa.

- Uma outra sugestão para melhorar a qualidade e confiabilidade dos registros é o próprio CNPq mobilizar os bolsistas de Iniciação Científica para verificação e atualização dos dados de seus orientadores. Os bolsistas teriam mais essa atribuição que também contribui para o aprendizado de aspectos da documentação e da comunicação científica na prática.[167]

Vale ressaltar a precisão na descrição dos autores. Graças ao uso de uma base de autoridade não foi encontrado nenhum tipo de erro nos nomes dos mesmos. Utilizando-se deste mesmo conceito, a base CL poderia se associar a uma base de autoridade de nomes de periódicos, como por exemplo a base do CCN (Catálogo Coletivo Nacional de Publicações Seriadas), mantida e disponibilizada pelo IBICT. A base do CCN poderia ser utilizada para validar ou rejeitar títulos de periódicos no momento em que o pesquisador dá entrada na base. Essa validação poderia ser facilmente implementada através do uso do ISSN.

O ISSN - Número Internacional Normalizado para Publicações Seriadas (*International Standard Serial Number*) é o identificador aceito internacionalmente para individualizar o título de uma publicação seriada, tornando-o único e definitivo. Seu uso é definido pela norma técnica internacional da *International Standards Organization* ISO 3297. O ISSN é operacionalizado por uma rede internacional, e no Brasil o Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT atua como Centro Nacional dessa

[167] Devo esta sugestão à generosidade do Prof. Luc Quonian, Diretor do CENDOTEC, durante reunião com minha orientadora para discutir o uso dos dados do CV Lattes em um projeto conduzido pelo IBICT em parceria com o Tecpar e o próprio Cendotec.

rede. O ISSN identifica o título de uma publicação seriada (jornais, revistas, anuários, relatórios, monografias seriadas, etc) em circulação, futuras (pré-publicações) e encerradas, em qualquer idioma ou suporte físico utilizado (IBICT, www.ibict.br).

Portanto, feitas essas considerações, elas reforçam o que se pretendia demonstrar, ou seja, os valores fornecidos pelo sistema Lattes apresentam um elevado grau de imprecisão devido aos fatores já anteriormente mencionados, comprometendo a finalidade da Plataforma Lattes como um sistema voltado para a construção de indicadores de C&T. Tal constatação pode explicar o porquê do próprio MCT ainda recorrer às bases internacionais como as do ISI (Institute for Scientific Information) para apresentar indicadores da produção bibliográfica nacional mesmo admitindo que parte substancial dos artigos produzidos no país é publicada em periódicos não indexados pela base de dados do ISI. Além das inúmeras deficiências que a base do ISI apresenta quando utilizada como fonte primária para construção de indicadores da produtividade científica ela subestima a produção científica dos países em desenvolvimento, incluindo aí o Brasil. Vale lembrar que esta base compila referências de cerca de 8 mil periódicos internacionais sendo que, destes, apenas cerca de 15 periódicos são brasileiros. Tais constatações levam, inevitavelmente, a indicadores com valores imprecisos e irreais, subestima a produção científica brasileira, comete injustiças com pesquisadores e instituições e desprestigia os periódicos nacionais. Daí decorre a enorme importância que a base Currículo Lattes adquire, sendo o único substituto a altura da base do ISI, como fonte primária de dados da produção científica brasileira para construção de indicadores de C&T. Entretanto, se por um lado os aspectos tecnológicos de armazenamento e disponibilização dos dados da base CL estão num patamar bem avançado graças a infra-estrutura existente e aos investimentos em TI proporcionados pelo CNPq, por outro lado, para que a base CL venha a produzir indicadores precisos e confiáveis faz-se necessário maior investimento em estudos de desenvolvimento de metodologias visando melhorar a qualidade dos dados nela contidos. Neste sentido, deve-se destacar o projeto SciELO, uma das iniciativas mais promissoras para a implantação de uma biblioteca virtual de revistas científicas brasileiras no formato eletrônico. Um dos objetivos previstos na Metodologia Scielo é a aprimorar o controle, a visibilidade e a avaliação da literatura científica brasileira.

A solução metodológica e tecnológica que o SciELO propicia poderá reparar duas deficiências que ocorrem com a base CL. A primeira é a duplicidade na contagem da produção bibliográfica. Isto é, quando, por exemplo, uma referência contém 4 autores, essa mesma referência será contabilizada quatro vezes nas estatísticas de produção científica disponibilizadas pelo sistema Lattes. A segunda falha se refere ao fato de que a simples associação do ano de publicação contido na referência com o período de permanência do autor em uma dada instituição não caracteriza necessariamente que aquela referência bibliográfica seja produção daquela instituição.

Derivar indicadores de vinculação institucional de autores é algo permitido a partir do registro, nas bases de dados bibliográficas, da procedência institucional dos autores de um dado trabalho científico. Cabe destacar que esse dado é geralmente incluído no próprio trabalho, o que permite que cada registro bibliográfico fique associado ao registro da instituição do autor(es). Como a base Lattes não é de natureza bibliográfica, não há esse campo para registro da vinculação institucional do autor associado a cada artigo. Como alternativa, a base CL estabelece essa associação partindo da suposição de que a correlação pode ser feita entre a data de publicação do trabalho e o período de permanência do autor em uma dada instituição. O uso de tal alternativa não estaria levando em consideração os seguintes aspectos:

- a instituição de um autor nem sempre é aquela que o emprega. Pode ser que ele tenha produzido o artigo enquanto esteve associado a um laboratório de pesquisa, por exemplo, fato que nem sempre acarreta seu registro no campo da experiência de trabalho;
- como há atraso na publicação, é possível que o trabalho tenha sido publicado quando o status institucional do autor já não era o mesmo presente quando da submissão do trabalho para publicação.

Portanto, o sucesso do projeto SciELO poderá trazer grande benefício para a qualidade dos dados da base CL pois o uso da linguagem XML associada a servidores de enlace irá permitir a associação direta e inequívoca da referência bibliográfica armazenada na base CL ao objeto digital referenciado, seja ele texto, imagem ou vídeo armazenado na base SciELO.

BIBLIOGRAFIA

ABOUT OCLC. **Online Computer Library Center**. Disponível em: http://www.oclc.org/about/ . Acesso em: mar. 2003.

ACCOMAZZI, A. *et al*. **The ADS bibliographic reference resolver**. San Francisco: Astronomical Society of the Pacific, 1999. Disponível em: http://monet.ncsa.uiuc.edu/adass98/Proceedings/accomazzia/accomazzia.html. Acesso em: nov. 2002.

A DESCRIPTION of database labels. **CIQM Database Labels**. Disponível em: http://www.la-hq.org.uk/liaison/ciqm/ciqmlbl2.html. Acesso em: jul. 2001.

ALMEIDA, M. B. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 5-13, maio/ago 2002.

APRESENTAÇÃO / histórico. **Indicadores de C&T**. Disponível em: http://www.mct.gov.br/estat/ascavpp/portugues/menu1page.htm. Acesso em: jul. 2003.

ARMSTRONG, C. Metadata, PICS and quality. **Ariadne**, v. 9, maio 1997. Disponível em: http://www.ariadne.ac.uk/issue9/pics/. Acesso em: dez. 2002.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TECNICAS. **Informação e documentação – referências – elaboração: NBR 6023**. Rio de Janeiro, ago. 2000. 22p.

AUN, M. P. A construção de políticas nacional e supranacional de informação: desafio para os estados nacionais e blocos regionais. **Ciência da Informação**, Brasília, v. 28, n. 2, 1999. Disponível em: http://www.ibict.br/cionline/280299/28029903.htm. Acesso em:nov. 2002.

BAX, M. P. Introdução às linguagens de marcas. **Ciência da Informação**, Brasília, v. 30, n. 1, p. 32-38, jan/abr 2001.

BARRÉ, R.; ESTERLE, L.; CHARLET, V. **Science and governance: the case of France**. Paris: OST, 2000. 21 p. Disponível em: http://www.obs-ost.fr/. Acesso em nov. 2002.

BATTAGLIA, M. G. B. **Análise sistêmico documental e proposta de um sistema de informação em C&T para a FINEP**. Rio de Janeiro: UFRJ, Escola de Comunicação – CNPq/IBICT, 1992. 112p. Dissertação. (Mestrado em Ciência da Informação).

BIREME. **Critérios de seleção de periódicos para a base LILACS**. São Paulo: BIREME, 2000. Disponível em: http://www.bireme.br/. Acesso em: nov. 2002.

BIREME. **Manual de descrição bibliográfica**. 4ª. ed., São Paulo: BIREME, 2000. 49p. Disponível em: http://www.bireme.br/. Acesso em: nov. 2002.

BOWEN, P. L.; FUHER, D. A. GUESS, F. M. Continuously improving data quality in persistent databases. **Data Quality**, Alexandria, EUA, v. 4, n. 1, set. 1998. Disponível em: http://www.dataquality.com/998bowen.htm. Acesso em: dez. 2002.

BRASIL: Esforços nacionais em C&T e dispêndios nacionais brutos em P&D. Indicadores de C&T. Disponível em: http://www.mct.gov.br/. Acesso em: nov. 2002.

BRICKLEY, D. *et al*. Recommendations on implementation of quality ratings in an RDF environment. In: **Project RE 4004 (RE): DESIRE II – Development of a european service for information on research and education II**. Deliverable 3.1, dez. 1998, 46p. Disponível em: http://www.desire.org/html/research/deliverables/D3.1/qualratings/doc0000.htm. Acesso em: nov. 2002.

CAMERON, R. D. A universal citation database as a catalyst for reform in scholarly communication. **First Monday**, v. 2, n.4, abr. 1997. Disponível em: http://www.firstmonday.dk/issues/issue2_4/cameron/index.html. Acesso em: nov. 2002.

CAMERON, R. D.; TATU, S. G. **Bibliographic protocol level 1: link resolution and metapage retrieval**. Internet Engineering Task Force (IETF), 2000. Disponível em: http://www.cs.sfu.ca/~cameron/bibp-revised.html. Acesso em: dez. 2002.

CENDÓN, B. V. Bases de dados de informação para negócios. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 30-43, maio/ago 2002.

CNPq. Plataforma Lattes. Disponível em: http://lattes.cnpq.br/. Acesso em fev. 2003.

CPRM. **Referências e citações bibliográficas: guia prático com exemplos em geociências**. Rio de Janeiro: CPRM/DIDOTE, 2001. 28p.

CUENCA, A. M. B. *et al*. Capacitação no uso das bases Medline e Lilacs: avaliação de conteúdo, estrutura e metodologia. **Ciência da Informação**, Brasília, v. 28, n. 3, p. 340-346, set/dez 1999.

DIAS, C. A. Portal corporativo: conceitos e características. **Ciência da Informação**, Brasília, v. 30, n.1, p. 50-60, jan/abr 2001.

DVIR, R.; EVANS, S. **A TQM aproach to the improvement of information quality**. Proceedings of the 1996 conference on Information Quality, MIT.

ESPANHA. Ministerio de Ciencia y Tecnologia. **Indicadores del sistema español de ciencia y tecnologia**. Madrid, 2000. 35 p.

FIRST Monday interviews: Cybrarian Reva Basch explores information and its uses in cyberspace. **First Monday**, v. 1, n. 4, out. 1996. Disponível em: http://firstmonday.org/issues/issue4/interview/index.html. Acesso em: jul. 2001.

FURNIVAL, A. C. A participação dos usuários no desenvolvimento de sistemas de informação. **Ciência da Informação**, Brasília, v. 25, n. 2, p. 1-13, 1995.

GALE/ALISE bibliographic instruction support program. Farmington Hills: Gale, 2001. Disponível em: http://www.galegroup.com/pdf/customer_service/alise.pdf. Acesso em: dez. 2002.

GUIA de implantação de sites SciELO. **Scientific Eletronic Library Online - SciELO**. Disponível em: http://www.scielo.org/guia_implantacion_pt.html . Acesso em: nov. 2002.

GUIMARÃES, R. **Avaliação e fomento de C&T no Brasil: propostas para os anos 90**. Brasília: MCT/CNPq, 1994. 178p.

GUIMARÃES, R. Diretório dos Grupos de Pesquisa: Apresentação. Disponível em: http://lattes.cnpq.br/diretorio. Acesso em: mar. 2001.

HEEMANN, V. **Avaliação ergonômica de interfaces de bases de dados por meio de *checklist* especializado**. Orientador: Walter de Abreu Cybis. Florianópolis: UFSC, 1997. Dissertação. (Mestrado em Engenharia da Produção). Disponível em: http://www.eps.ufsc.br/disserta97/heemann/. Acesso em: nov. 2002.

HERNÁNDEZ-ORALLO, J. **Knowledge discovery in databases and data quality**. 1999. Disponível em: http://www.dsic.upv.es/~jorallo/KDD/KDD.html. Acesso em: dez. 2002.

HOFMAN, P. *et al*. Specification for resource description methods Part 2: Selection criteria for quality controlled information gateways. In: **Project RE 1004 (RE): DESIRE – Development of a european service for information on research and education**. Deliverable D3.22, mar. 1996, 90p. Disponível em: http://www.ukoln.ac.uk/metadata/desire/quality/. Acesso em: nov. 2002.

KIELGAST, S.; HUBBARD, B. A. Valor agregado à informação – da teoria a prática. **Ciência da Informação**, Brasília, v.26, n. 3, 1997. Disponível em: http://www.ibict.br/cionline/260397/26039706.htm. Acesso em: nov. 2002.

KONDO, E. K. Desenvolvendo indicadores estratégicos em ciência e tecnologia: as principais questões. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 128-133, maio/ago 1998.

KUNY, T. Filtering Internet content: PICS, labels and filters. **Network Notes**, Ottawa: National Library of Canada, v. 53, mar. 1998. Disponível em: http://www.nlc-bnc.ca/9/1/p1-252-e.html. Acesso em: dez. 2002.

LASSILA, O.; SWICK, R. R. (Ed.). **Resource description framework (RDF) model and syntax specification**. W3C (MIT, INRIA, Keio), 1999. Disponível em: http://www.w3.org/TR/REC-rdf-syntax/. Acesso em: dez. 2002.

LASTRES, H. M. M. Dilemas da política científica e tecnológica. **Ciência da Informação**, Brasília, v. 24, n. 2, 1995.

LASTRES, H. M. M. Informação e conhecimento na nova ordem mundial. **Ciência da Informação**, Brasília, v. 28, n. 1, 1999. Disponível em: http://www.ibict.br/cionline/280199/28019910.htm. Acesso em:nov. 2002.

MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 134-140, maio/ago 1998.

MARCONDES, C. H. Representação e economia da informação. **Ciência da Informação**, Brasília, v. 30, n. 1, p. 61-70, jan/abr 2001.

MARCONDES, C. H.; SAYÃO, F. F. 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.

MATTHEWS, J. The value of information in library catalogs. **Information Outlook**, Washington: Special Libraries Association - SLA, jul. 2000. Disponível em: http://www.sla.org/pubs/serial/io/2000/jul00/jmatthews.html. Acesso em: nov. 2002.

METODOLOGIA e conceitos. Indicadores de C&T. Disponível em: http://www.mct.gov.br/estat/ascavpp/portugues/menu9page.htm. Acesso em: jul. 2003.

MIRANDA, D. B. O periódico científico como veículo de comunicação: uma revisão de literatura. **Ciência da Informação**, Brasília, v. 25, n. 3, 1996.

MORESI, E. A D. Delineando o valor do sistema de informação de uma organização. **Ciência da Informação**, Brasília, v. 29, n. 1, p. 14-24, jan/abr 2000.

MOURA, L. R. Informação: a essência da qualidade. **Ciência da Informação**, Brasília, v. 25, n. 1, 1995.

MUSTAR, P. **Les chifres clés de la science & de la technologie**. Ed. 1998-1999, Paris: OST, 1998. 111p.

OBSERVATOIRE DES SIENCES E DES TECHNIQUES. **Science & tecnologie: indicateurs 1998**. Paris: Econômica, 1998. 551p.

OCLC. Introduction. In: **Bibliographic formats and standards guide**. Dublin, EUA: OCLC Online Computer Library Center, 2002. Disponível em: http://www.oclc.org/bibformats/en/introduction/. Acesso em: dez. 2002.

OCLC. Quality Assurance. In: **Bibliographic formats and standards guide**. Dublin, EUA: OCLC Online Computer Library Center, 2002. Disponível em: http://www.oclc.org/bibformats/en/quality/. Acesso em: dez. 2002.

O'NEIL, E. T.; VIZINE-GOETZ, D. Quality control in on-line databases. In: WILLIAMS, M. E., ed. **Annual review of information science and technology (ARIST)**. New Jersey: Elsevier-ASIS, v. 23, 1988. p. 125-156.

PACHECO, R. C. 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. L. *et al*. SciELO: uma metodologia para publicação eletrônica. **Ciência da Informação**, Brasília, v. 27, n. 2, 1998. Disponível em: http://www.ibict.br/cionline/270298/27029802.htm. Acesso em: nov. 2002.

PEDRINI, A. G. **O cientista e os métodos de avaliação de seu desempenho: estudo de sua adequação no contexto brasileiro**. Orientador: Rosali Fernandez de Souza. Rio de Janeiro: UFRJ, Escola de Comunicação – CNPq/IBICT, 1999. 442p. Tese. (Doutorado em Ciência da Informação).

PEREIRA, Maria de Nazaré Freitas. **Por uma Economia do Conhecimento**: Avaliação de Bases de Dados Nacionais para a Produção de Indicadores de C&T (Ciência e Tecnologia). Relatório Parcial (Avaliação de qualidade de bases de dados bibliográficas). Rio de Janeiro, julho/2001. Processo 520416/93-7 (NV).

PEREIRA, M. N. F. *et al*. Bases de dados na economia do conhecimento: a questão da qualidade. **Ciência da Informação**, Brasília, v.28, n. 2, 1999. Disponível em: http://www.ibict.br/cionline/280299/28029913.htm. Acesso em:nov. 2002.

PESSANHA, C. Critérios editoriais de avaliação científica: notas para discussão. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 226-229, maio/ago 1998.

PINTO, M. M. N. **Indicadores de P&D do setor produtivo no Brasil:** situação, necessidades e perspectivas. Orientador: Paulo César Gonçalves Egler. Brasília: Universidade de Brasília, Centro de Desenvolvimento Sustentável, 2000. 74p. Dissertação. (Mestrado em Desenvolvimento Sustentável).

PLATAFORM for Internet content selection (PICS). W3C. Disponível em: http://www.w3.org/PICS/. Acesso em: dez. 2002.

PRODUÇÃO científica. Indicadores de ciência e tecnologia em São Paulo. Disponível em: http://www.fapesp.br/indct/pag89.htm. Acesso em: abr. 2001.

PRODUÇÃO científica. Metodologia e conceitos. **Indicadores de C&T**. Disponível em: http://www.mct.gov.br/estat/ascavpp/6_Producao_Cientifica /notas/txt_prod_cient.htm. Acesso em: out. 2001.

PROGRAMA de catalogação cooperativa (PCC). Preparado pela equipe da Library of Congress Hispanic Reading Room; editado pela equipe de Catalogação

Cooperativa, março 1999; traduzido para o português sob a responsabilidade do Departamento Técnico do Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBi/USP), São Paulo, Brasil, nov. 1999. Disponível em: http://www.loc.gov/catdir/pcc/pccpor.html. Acesso em: nov. 2002.

QUALITY on the Internet. **db-Qual**, v. 2, n. 1, jan. 1997. Disponível em: http://www.la-hq.org.uk/liaison/ciqm/dbq_3_4.html. Acesso em: dez. 2002.

RIOS, R.; SANTANA, P. H. A El espacio virtual de intercambio de información sobre recursos humanos em ciencia y tecnología de América Latina e Caribe: Del CV Lattes al CvLAC. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 42-47, set/dez 2001.

SANCHO, R. Indicadores bibliométricos utilizados em la evaluacion de la ciência y la tecnologia, revision bibliográfica. **Revista Española de Documentación Cientifica**, Madrid, v. 13, n. 3-4, p. 842 –865, 1990.

SANTANA, P. H. A. *et al*. Servidor de enlaces: motivação e metodologia. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 48-55, set/dez 2001.

SAYÃO, L. F. Bases de dados: a metáfora da memória científica. **Ciência da Informação**, Brasília, v. 25, n. 3, 1996.

SAYÃO, L. F. Bases de dados e suas qualidades. In: LUBISCO, N.; BRANDÃO, L. (Ed.). **Informação e Informática**. Salvador: EDUFBA, 2000.

SILVA, G. L. A política da União Européia no domínio da informação científico-tecnológica. **Ciência da Informação**, Brasília, v.26, n. 1, 1997. Disponível em: http://www.ibict.br/cionline/260197/26019709.htm. Acesso em: nov. 2002.

SMITH, A. **Criteria for evaluation of Internet information resources**. Canberra: Information Quality WWW Virtual Library, 1997. Disponível em: http://www2.vuw.ac.nz/staff/alastair_smith/evaln/. Acesso em: dez. 2002.

SPINAK, E. Indicadores cienciométricos. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 141-148, maio/ago 1998.

STREHL, L. Avaliação da consistência da indexação realizada em uma biblioteca universitária de artes. **Ciência da Informação**, Brasília, v. 27, n.3, p. 329-335, set/dez 1998.

STUMPF, I. R. C. Passado e futuro das revistas científicas. **Ciência da Informação**, Brasília, v. 25, n. 3, 1996.

STUMPF, I. R. C. Reflexões sobre as revistas brasileiras. **InTexto**, Porto Alegre, v.1, n. 3, 1998. Disponível em: http://www.ilea.ufrgs.br/intexto/. Acesso em: set. 2001.

TARGINO M. 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.

TESTA, J. A base de dados ISI e seu processo de seleção de revistas. **Ciência da Informação**, Brasília,v. 27, n. 2, p. 233-235, maio/ago 1998.

TWIDALE, M. B.; MARTY, P. F. An investigation of data quality and collaboration. **Technical Report ISRN UIUCLIS--1999/9+CSCW**, 1999. Disponível em: http://www.lis.uiuc.edu/~twidale/pubs/dq.html. Acesso em: dez. 2002.

VALENTIM, M. L. P. A indústria da informação e os produtores de bases de dados em C&T. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 7, n. 1, p. 23-37, jan/jun 2002.

VALSS, V. M. O gerenciamento dos documentos do sistema da qualidade. **Ciência da Informação**, Brasília, v. 25, n. 2, 1995.

VELHO, L. Indicadores científicos: aspectos teóricos y metodológicos. In: MARTINEZ, E. (ed.). **Ciencia, tecnologia y desarrollo: interrelaciones teóricas y metodológicas**, Caracas: Nueva Sociedad, 1994. p. 307-348.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: what data quality means to data consumers. **Journal of Management Information Systems**, v. 12, n. 4, p. 5-33,1996.

ANEXOS