

A Importância do papel do profissional da ciência da informação nos processos de recuperação de conteúdos digitais estruturados

Ricardo César Gonçalves Sant'Ana

Como citar: SANT'ANA, R. C. G. A Importância do papel do profissional da ciência da informação nos processos de recuperação de conteúdos digitais estruturados. *In:* FUGITA, M. S. L.; GUIMARÃES, J. A. C. **Ensino e Pesquisa em Biblioteconomia no Brasil: a emergência de um novo olhar.** Marília: Ed FUNDEPE, 2008. p145-154



All the contents of this work, except where otherwise noted, is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported.

Todo o conteúdo deste trabalho, exceto quando houver ressalva, é publicado sob a licença Creative Commons Atribuição - Uso Não Comercial - Partilha nos Mesmos Termos 3.0 Não adaptada.

Todo el contenido de esta obra, excepto donde se indique lo contrario, está bajo licencia de la licencia Creative Commons Reconocimiento-NoComercial-CompartirIgual 3.0 Unported.

A importância do papel do profissional da ciência da informação nos processos de recuperação de conteúdos digitais estruturados

Ricardo César Gonçalves Sant'Ana

A idéia para elaboração deste texto surgiu da necessidade de análise constante sobre que conteúdo deve conter as disciplinas de Introdução à Ciência da Computação e Introdução à Editoração ministradas em cursos de Ciência da Informação. Disciplinas estas que trabalham parte das competências a serem desenvolvidas pelos futuros profissionais da informação em sua interação com as tecnologias da informação.

Para responder questões como: Quais temas abordar? Com que profundidade? Somos levados, primeiramente, a responder outras questões como: qual é o papel do profissional da ciência da informação no processo de recuperação de conteúdos digitais? Qual deve ser seu nível de interação com a tecnologia da informação? Quais competências são necessárias e precisam ser trabalhadas durante sua formação acadêmica?

Este texto não se propõe a apresentar respostas, mas sim apresentar elementos que subsidiem uma futura discussão mais detalhada sobre o assunto, trazendo considerações resultantes de reflexão sobre aulas ministradas nos cursos de Arquivologia e Biblioteconomia, e de discussões e troca de idéias com os alunos e colegas docentes.

Para entendermos de forma mais clara o papel do profissional da Ciência da Informação na interação com as tecnologias da informação, precisamos de uma análise prévia do processo de uso destas tecnologias no acesso aos conteúdos digitais.

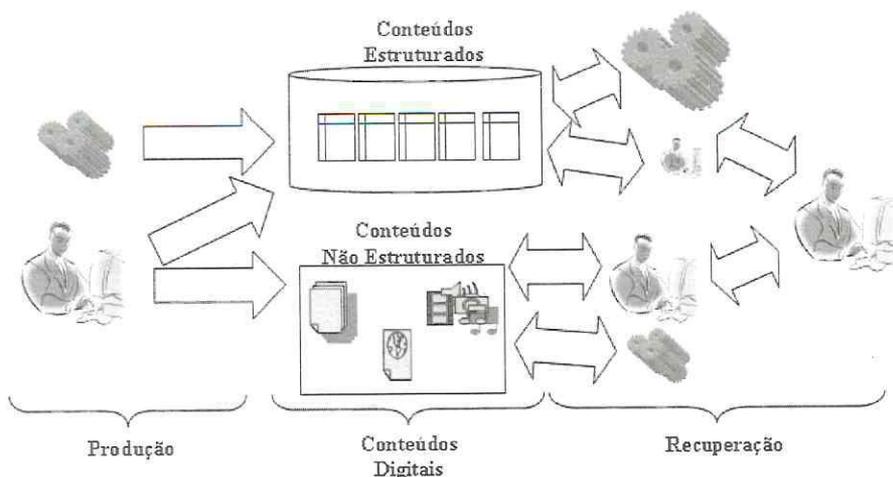


Figura 1: Conteúdos digitais estruturados e não estruturados.

Com a adoção maciça das novas tecnologias da informação e comunicação, o volume de informações armazenadas e disponíveis para acesso vem crescendo de forma exponencial, mas transformar este volume em valor agregado para os usuários requer um processo de recuperação cada vez mais eficiente. Todo este conteúdo digital não pode ser tratado de forma única quando se tem como foco o processo de recuperação. Uma das divisões que podem ser feitas é com relação à estruturação destes conteúdos. Assim, temos, por um lado, os conteúdos que são gerados por processos automatizados de funções operacionais de organizações em seus mais diversos sistemas. Estes conteúdos têm, como função primária, atender as necessidades de sistemas que devem realizar tarefas previamente definidas e bem estruturadas.

Como exemplo, podemos pensar em um sistema de faturamento (HABERKORN, 1999) que visa automatizar o processo das vendas realizadas aos clientes, emitindo, assim, de forma eficiente e eficaz, notas fiscais e, eventualmente, faturas, alimentando ainda sistemas como o de contas a receber, contábil, fluxo de caixa, entre outros. Portanto, seu objetivo primário não é atender as consultas realizadas pelos usuários que desejam recuperar informações gerenciais ou consolidadas. São aplicações, interfaces e bases de dados projetadas para o dia a dia dos usuários envolvidos com a 'operação' do processo de vendas. Para atender a estes objetivos, os dados são armazenados nestas bases em estruturas previamente definidas e agrupadas em conjuntos maiores com outros elementos similares que contêm a mesma estrutura. Tomando como exemplo um hipotético sistema de faturamento podemos propor uma estrutura simplificada em que estes dados poderiam estar organizados, conforme ilustrado na Figura 2.

Podemos verificar, assim, que todos os dados armazenados terão seu conteúdo estruturado de forma simples e direta, ou seja, quando se deseja saber o endereço de um deter-

minado cliente basta acessar a tabela “Cliente” e, então, localizar o cliente desejado a partir de alguma informação prévia como código ou nome e, então, obtendo o conteúdo do atributo “endereço” recuperamos a informação desejada.

A alimentação destas tabelas é realizada através de interfaces disponibilizadas a usuários, ou grupos de usuários específicos ou pelo processo de automatização do próprio fluxo de informação e integração de sistemas dentro da organização. Um exemplo de geração automática de informação, seguindo ainda o nosso exemplo de sistema de faturamento, seria a alimentação de um sistema de contas a receber que receberia a informação do faturamento, possibilitando, assim, o acompanhamento do recebimento dos valores gerados pelas vendas.

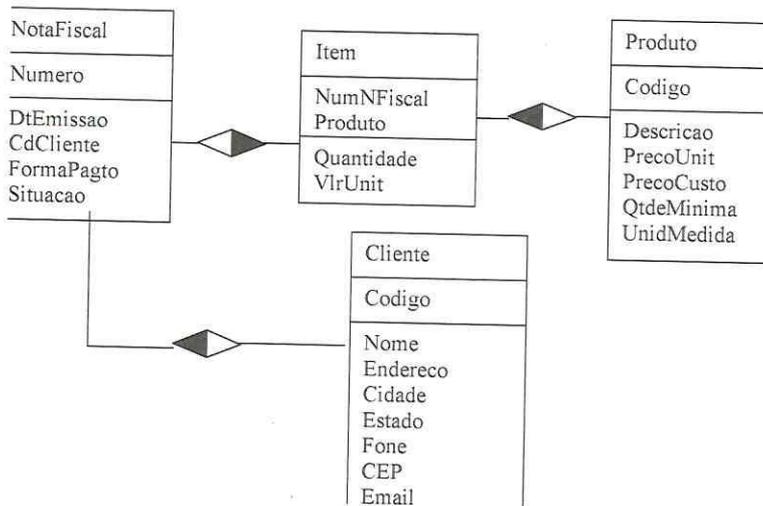


Figura 2: Modelo de dados simplificado de um sistema de faturamento.

O processo de recuperação de conteúdos digitais estruturados é realizado com uma interferência muito maior de recursos da própria tecnologia da informação, que permite a elaboração de algoritmos e predefinições de caminhos para que a localização de informações desejadas seja simples e eficiente. No entanto, mesmo no caso de informações estruturadas, algumas necessidades não têm como ser previstas, exigindo recursos complementares.

Por outro lado, temos os conteúdos não estruturados que são gerados em formato de texto, como, por exemplo, uma notícia, um artigo, o registro de um diálogo, comunicação escrita, um contrato, ou qualquer outro documento que não siga uma formatação fixa e que, portanto, necessita de uma interpretação para que sejam criados mecanismos de localização ou mesmo de identificação de cada objeto. Nesta categoria poderíamos também enquadrar as imagens, sons, animações ou qualquer outro conteúdo multimídia.

Para este tipo de objeto, a solução, que geralmente é utilizada para permitir a identificação e recuperação, consiste em criar um conjunto de dados sobre cada objeto que siga uma estrutura previamente estabelecida e que, assim, possa ser utilizado pela própria tecnologia no processo de recuperação. Estes dados adicionais gerados sobre os objetos de informação são os metadados, ou seja, dados sobre os dados.

A alimentação destes metadados, no entanto, ainda necessita de uma interação humana mais ativa, já que depende de um processo de interpretação e de elaboração que as máquinas ainda não são capazes de resolver. A interação humana também se faz presente na definição de qual o padrão de metadados será adotado, definição esta que depende do tipo de objeto e do tipo de aplicação que irá utilizá-lo.

Recuperação de conteúdos digitais estruturados

Os conteúdos digitais estruturados possibilitam uma grande variedade de mecanismos de recuperação e de tratamento, variedade esta ampliada pelos próprios recursos disponibilizados pelas tecnologias da informação (Figura 3). Como primeira opção podemos citar a possibilidade de localização direta disponibilizadas pelas interfaces de operação das próprias aplicações que geram estas informações e que são de grande utilidade para buscas mais simples e que estão baseadas em necessidades mais simples, como, por exemplo, a localização dos dados de um determinado cliente a partir do seu código ou do seu nome, ou da apresentação dos dados de uma determinada nota fiscal a partir do seu número, e assim por diante.

Em contra partida à sua facilidade de utilização, temos a limitação de opções que oferece, sendo, assim, de uso mais corrente para os próprios usuários do processo operacional do sistema.

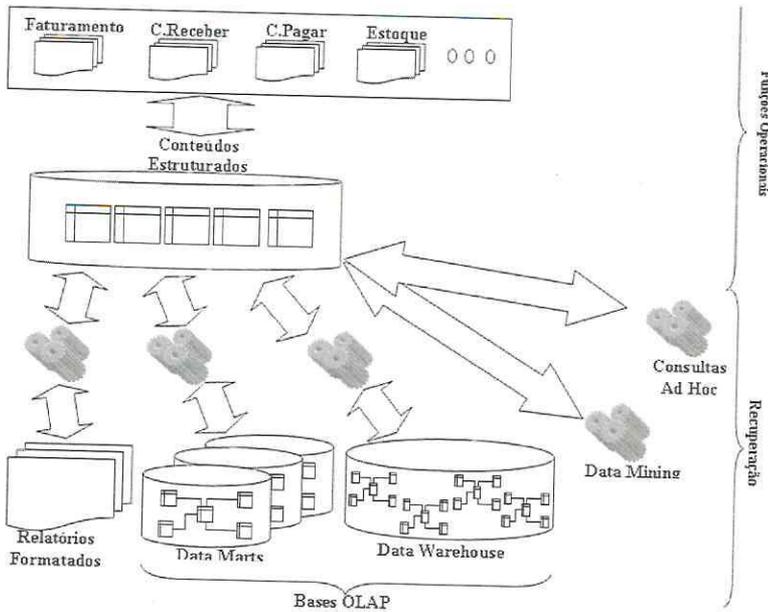


Figura 3: Processos de recuperação de conteúdos digitalizados.

Outra possibilidade se origina, geralmente, a partir do trabalho de analistas de sistemas ou de analistas de negócios que identificam as possíveis necessidades de informação que cada sistema pode vir a gerar, pelo menos as mais correntes, e projetam relatórios que podem ser emitidos diretamente pelo sistema ou após processamentos prévios, gerando relatórios formatados (Figura 3) que podem, inclusive, ser visualizados diretamente na tela do próprio aplicativo do sistema.

Assim como no caso anterior, esta opção tem como principal característica a limitação de opções de acesso as informações da base, já que dependem de uma previsão dos analistas. Outra limitação é inerente a própria arquitetura em que a base de dados foi projetada, já que a mesma tem como objetivo primário a performance do processo operacional da organização e não atender a consultas de seu conteúdo.

Para contornar esta limitação pode-se optar pela adoção de bases de dados replicadas, ou seja, geradas a partir das bases de dados principais, mas com uma arquitetura que tenha como principal objetivo o atendimento às necessidades de recuperação de informações. Da distinção entre estas duas bases que passam a coexistir, define-se dois conceitos utilizados pela área de administração de empresas e pela área da ciência da computação: OLTP e OLAP (SINGH, 2001).

O primeiro, OLTP (*On-Line Transaction Processing*), ou processamento de transações em tempo real, identifica as soluções que têm como objetivo o registro e operacionalização de processos correntes da organização, como o nosso exemplo de sistema de faturamento, enquanto que, OLAP (*On-Line Analytical Processing*) ou processamento analítico em tempo real, cunhado pela primeira vez por Edward F Codd, identifica tecnologias projetadas exclusivamente para suprir as necessidades de obtenção de informações resultantes de consultas complexas de forma simples e eficiente.

Outra característica que diferencia estas duas tecnologias é a forma com que as bases são esvaziadas, ou seja, a forma como é definido o tempo em que os dados permanecerão disponíveis, sendo que o OLTP tem como base as exigências legais, fiscais e operacionais de necessidade de permanência em disponibilidade dos dados de cada transação enquanto que as bases OLAP tendem a manter períodos maiores em função da necessidade de análises históricas e de cunho comparativo entre períodos, tendo assim um horizonte de tempo bem mais amplo.

Uma base OLAP pode ser gerada a partir de dados de um determinado departamento, contendo os dados principais, que poderão ser consultados, já devidamente consolidados, e tratados de tal forma que possam agregar valor aos conteúdos recuperados. Estas bases de dados são denominadas Data Marts³ (Figura 3) e são construídas com um formato que traz, ainda, possíveis dados complementares que permitem a parametrização das consultas, ou seja, a liberdade de escolha do usuário para definir parâmetros que formatam o resultado que desejam obter sob diversos pontos de vista, ou dimensões.

Assim, temos uma arquitetura diferente para o modelo de dados que é adotada pelos sistemas OLTP. Entre as possibilidades para modelar estas estruturas (tabelas), que irão compor as bases de dados OLAP, temos o modelo STAR, Figura 4, que é composto por uma tabela principal denominada “Fato” e tabelas complementares que permitem a visualização sob diferentes pontos de vista, que são as tabelas de “Dimensões”, destacando-se entre elas a tabela da dimensão tempo.

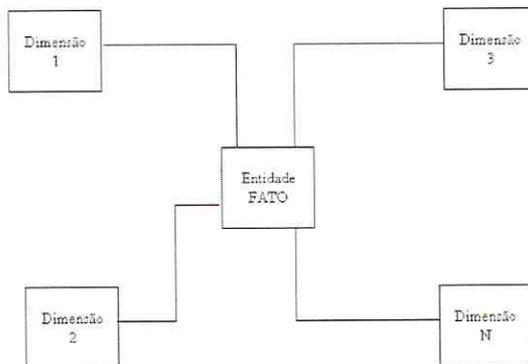


Figura 4: Modelo de dados STAR para uma aplicação OLAP.

A adoção de Data Marts em diversos departamentos prepara a organização para projetos maiores e mais completos que possam inclusive otimizar os recursos investidos nestas soluções como a aquisição de equipamentos adicionais e aplicativos específicos.

Um próximo passo, então, pode ser a implementação de um projeto OLAP com a utilização de várias bases de dados como as utilizadas pelos Data Marts, porém unificadas em um projeto único, que passa a ser definido como sendo um Data Warehouse (INMON, 1996; KIMBALL, 1996) e que tem como principais características seu alto custo e forte impacto sobre a organização como um todo, em função de sua profunda dependência de interligação de dados oriundos de bases de dados de diversos departamentos e sistemas.

Outro ponto que merece atenção no processo de implementação de um Data Warehouse, está ligado às questões ligadas à utilização, em conjunto, de conceitos que muitas vezes são definidos, dentro da mesma organização, de forma diferente. Por exemplo, a definição de “faturamento líquido” pode ter uma composição de variáveis para o departamento de vendas e outra diferente para o departamento fiscal.

Outra necessidade a ser contornada é a identificação de dados novos e que possam ser úteis, através da busca e análise de padrões relevantes nos grandes volumes de dados gerados e registrados nas bases de dados das organizações. Este conjunto de tecnologias é definido como Data Mining (Figura 3) e também pode ser de grande valor para os processos de recuperação em conteúdos digitais estruturados.

Porém, nem todos estes recursos juntos conseguem atender a todas as possibilidades de necessidades de buscas por informações que os usuários, principalmente aqueles envolvidos nos processos decisórios, podem apresentar.

Para suprir as necessidades esporádicas e específicas que não são passíveis de serem previstas ou não têm uma periodicidade que justifique investimento específico, as soluções de consulta *Ad Hoc* (Figura 3) podem representar uma boa alternativa. Apesar de não apresentarem bom desempenho, proporcionam tamanho grau de flexibilidade, que permitem que os próprios usuários, muitas vezes com ajuda de suporte técnico, possam realizar suas consultas com alto grau de parametrização.

Conteúdos digitais estruturados e o profissional da Ciência da Informação

Analisando, assim, o contexto dos conteúdos digitais estruturados e os diversos recursos envolvidos nos processos de busca e recuperação, percebe-se que o envolvimento maior ainda cabe aos profissionais mais capacitados na questão de uso e interação com as tecnologias da informação. Qual o papel que pode, então, o novo profissional da ciência da

informação ocupar neste cenário? Um profissional que “distanciado do informático puro e duro dos anos setenta e oitenta e, entretanto, diferenciado do gestor de recursos informáticos dos anos noventa [...], na medida em que tende a combinar conhecimentos e competências na área de gestão e planejamento estratégico, recursos humanos e marketing com diretrizes elementares no campo tecnológico de molde a intervir com credibilidade e êxito em domínios emergentes” do uso das TICs. (SILVA; RIBEIRO, 2002).

No que tange a interação mais direta com os recursos das Tecnologias da Informação, principalmente no que diz respeito aos processos de busca e recuperação, o profissional da CI pode atuar em diversos níveis, desde o controle e acompanhamento do uso e aproveitamento dos recursos para obtenção de informação dentro das organizações, respondendo de forma holística a questões tais como: qual o retorno sobre determinado investimento sobre uma ferramenta de apoio a processos de recuperação? Questões estas que geralmente apresentam respostas baseadas em elementos intangíveis e que, portanto, exigem uma análise mais completa e detalhada.

No tratamento de questões ligadas a atritos gerados por divergências ocasionadas pela sobreposição de sistemas conceituais, o papel do profissional da CI também pode ser de grande relevância, identificando o problema e propondo soluções, tanto em termos de processo como de possíveis alternativas de uso da tecnologia da informação.

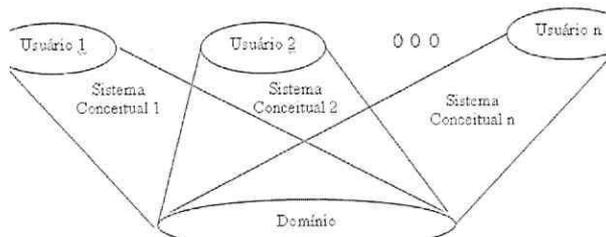


Figura 5: Usuários agindo em um mesmo domínio.

Fonte: Adaptado de: SHAW; GAINES, 1989.

A Figura 5 ilustra a sobreposição de sistemas conceituais de usuários diferentes sobre um mesmo domínio. Sendo que este domínio pode representar, por exemplo, um resultado obtido em um processo de recuperação, conforme já discutido neste texto.

A forte integração de sistemas, incrementada nos últimos anos com a adoção pelas organizações de uma solução única para todos os processos operacionais internos, tornou esta questão um dos fatores chave de sucesso na utilização e tratamento de conteúdos recuperados. Estas soluções únicas são denominadas ERP (*Enterprise Resource Planning*) e visam a unificação das tecnologias de informações aplicadas aos processos transacionais (OLTP).

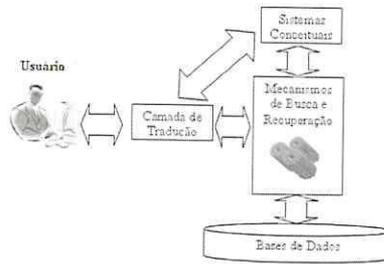


Figura 6: Criação de uma camada de tradução entre o usuário e o conteúdo recuperado.

Fonte: SANT'ANA, 2002.

Outra atuação importante que pode ser realizada pelo profissional de CI, ainda com relação a conteúdos digitais estruturados, é na questão ligada à definição e implementação de uma camada lógica de tradução (Figura 6) para o acesso aos dados recuperados, camada esta que permite um melhor entendimento das informações contidas em uma base de dados.

É através desta camada que denominações técnicas, utilizadas para identificar tabelas e seus atributos, podem ser convertidas em termos mais próximos dos utilizados pelos usuários finais. Pode, ainda, funcionar como uma alternativa para solução de conflitos de sobreposição de diferentes sistemas conceituais utilizados pelos usuários na interpretação de informações recuperadas.

As possibilidades de atuação do profissional da CI, no contexto dos conteúdos digitais estruturados, não param por aí, mas o escopo deste texto não permite um detalhamento de todas as possibilidades, restringindo-se à apresentação dos exemplos mais significativos.

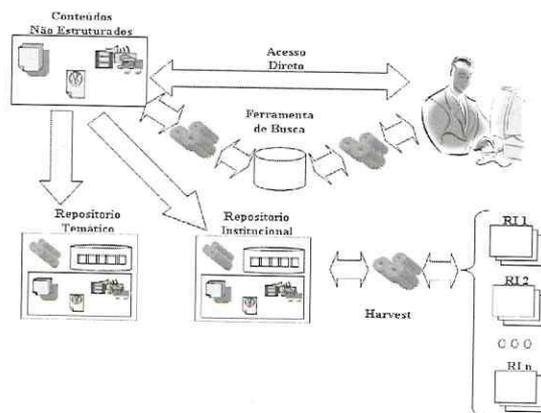


Figura 7: Exemplos de processos de armazenamento e recuperação de conteúdos não estruturados.

Cabe, ainda, destacar que, no âmbito dos conteúdos digitais não estruturados, o papel do profissional de CI é ainda mais importante e diferenciado, já que a interação humana na interpretação e identificação destes conteúdos ainda não pode ser suprida por processos automatizados.

Não menos importante, também, é a sua possível participação nas decisões sobre qual tecnologia adotar, como implementar, e quais políticas deverão ser seguidas com relação à implantação e uso destas tecnologias, ações estas, que têm características e peculiaridades, que fazem com que este profissional tenha grandes possibilidades de atuação.

Referências

- ABERKORN, E. **Teoria do ERP: enterprise resource planning**. 2. ed. São Paulo: Makron Books, 1999.
- INMON, W. H. **Building the data warehouse**. 2. ed. New York: John Wiley & Sons, 1996.
- KIMBALL, R. **The data warehouse toolkit: practical techniques for building dimensional data warehouses**. New York: John Wiley & Sons, 1996.
- SANT'ANA, R. **Unidades de conhecimento: fatores e métricas**. 2002. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2002.
- SHAW, M. L. G.; GAINES, B. R. **Comparing conceptual structures: consensus, conflict, correspondence and contrast**. Alberta: Knowledge Science Institute, University of Calgary, 1989.
- SILVA, A. M.; RIBEIRO, F. **Das ciências documentais à Ciência da Informação: ensaio epistemológico para um novo modelo curricular**. Porto: Edições Afrontamento, 2002.
- SINGH, H. **Data Warehouse: conceitos, tecnologias, implementação e gerenciamento**. São Paulo: Makron Books, 2001.