Representação de informações não estruturadas em ambientes corporativos: proposta de modelo conceitual

Sergio de Castro Martins[1][https://orcid.org/0000-0003-2406-6648], Carlos Henrique Marcondes[2][https://orcid.org/0000-0003-0929-8475]

[1] Departamento de Ciência da Informação, Universidade Federal Fluminense, Brasil.
smartins@id.uff.br
[2] Departamento de Ciência da Informação, Universidade Federal Fluminense, Brasil.
ch.marcondes@id.uff.br

Resumen. A pesquisa tem como propósito o estabelecimento de um modelo genérico de gestão da informação corporativo por meio de enriquecimento semântico de metadados. Pretende estender o modelo de Soergel para aplicação em contexto corporativo típico, considerando dados não estruturados em formato não somente textual, mas também multimídia. Serão expostos exemplos ou casos para aplicação do processo de destilação de dados e enriquecimento semântico de metadados, como processos complementares ao modelo de Soergel. Para o atingimento dos objetivos expostos, foi definido que a pesquisa será de caráter qualitativo e exploratório, sendo caracterizada como pesquisa-ação, na qual a pesquisa é orientada à busca de resultados específicos ou soluções inovadoras de problemas observados num ambiente empírico. As etapas previstas do modelo são Ambientação, Definição de tipos e fonte de dados, Destilação dos dados, Enriquecimento de metadados e Armazenamento. Como resultado, espera-se que o modelo estendido permita processar dados heterogêneos segundo os recortes estabelecidos e mediante os processamentos acima elencados, permitindo criação de valor na composição de documentos dinâmicos com agregações semânticas adicionados aos metadados.

Palabras clave: Organização e representação da informação, Enriquecimento semântico de metadados, Dados não estruturados, Modelo Conceitual.

Abstract. The research aims to establish a generic model of corporate information management through semantic enrichment of metadata. It intends to extend the Soergel's model for application in a typical corporate environment, considering unstructured data not only in textual format but also multimedia. Examples or cases will be presented for application of the process of data distillation and semantic enrichment of metadata, as complementary processes to Soergel's model. To achieve the stated objectives, it was defined that the research will be qualitative and exploratory, being characterized as “action research”, in which the research is oriented towards the search for specific results or innovative solutions of problems obeyed in an empirical environment. The expected steps in the model are Environment, Data Type and Data Source Definition, Data Distillation, Metadata Enrichment, and Storage. As a result, it is expected that the robust model will allow heterogeneous data processing according to the established cutouts and through the above-mentioned processes, allowing the creation of value

in the composition of dynamic documents with semantic aggregations added to the metadata.

Keywords: Information organization and representation, Semantic enrichment of metadata, Unstructured Data, Conceptual model.

1 Contextualização

Muito já tem sido apontado e discutido sobre a velocidade e impacto das alterações e rupturas pelas quais a sociedade moderna convive na atualidade, sobretudo pela utilização massiva das tecnologias da informação e comunicação (TICs). Após o advento da internet, pode-se afirmar que não somente mais uma nova era da informação teve início, mas também uma nova revolução industrial (Rifkin, 2012, Schwab, 2016). Tendo em conta que as corporações são empreendimentos comerciais responsáveis pelo provimento de bens e serviços requeridos não somente por indivíduos, mas também por qualquer tipo de coletividade humana, como governos e instituições, seu fluxo informacional apresenta-se como um ambiente altamente complexo. Praticamente toda aquisição de bem ou serviço gera dados e documentos necessários para sua concretização. Para além do setor comercial, também no setor industrial os processos produtivos requerem não somente matérias-primas, mas também insumos de dados e instruções para a produção (Swanson, 2012). No que se refere à economia de serviços, muitos bens digitais já começaram a substituir seus congêneres físicos como maior quantitativo volumétrico em vendas. Desde livros, filmes e discos, ainda que sejam comercializados fisicamente, suas versões em meio digital – e-books, streamings e mp3 – já possuem índices significativos de preferência do público usuário ou consumidor.

Particularmente no contexto corporativo, a influência dos dados na tomada de decisões estratégicas ou operacionais têm tido um destaque crescente, e praticamente nenhuma operação pode ser realizada sem a veiculação de carga de dados. Se há apenas algumas décadas a carga de dados podia ser suportada por *Data Warehouses* e bancos de dados tradicionais, na atualidade o volume, a variedade e a velocidade dos dados que circulam em uma companhia já não são suficientemente suportados pelos meios tradicionais de processamento da informação. Desta forma, novas metodologias de processamento de dados têm sido desenvolvidas nas últimas décadas, bem como novos sistemas de armazenamento. Adicionalmente a estes fatores, a própria característica dos dados corporativos representam um desafio: de acordo com Brocke e Simons (2014) e Marquesone (2015), mais de 80% dos dados são não-estruturados, tais como imagens, vídeos e áudios, além de textos diversos. Além disso, os dados provêm de variadas fontes internas e externas à organização, circulando por um ecossistema que Inmon (2001, 2008) define como *Corporate Information Factory* (CIF). O CIF consiste numa infraestrutura e cadeia de integração de vários elementos, como mundo exterior, aplicações, dados, metadados e tecnologias de processamento e armazenamento.

Proveniente de fontes diversas, dados e informações no complexo ambiente corporativo podem ser gerados por humanos e máquinas (Marquesone, 2015, Inmon, 2016). De modo geral, os conjuntos de dados no ambiente corporativo podem ser classificados de maneira levemente distinta.

Sob qualquer ótica, os dados são elementos essenciais ao funcionamento corporativo. Enquanto alguns dados e documentos são produzidos em ciclos de meses, outros são produzidos e consumidos em tempo real. Neste ambiente, o fluxo de dados internos veiculados entre departamentos correspondem a uma fração tão expressiva quanto dados provenientes ou enviados para fontes externas. Para o armazenamento e gerenciamento da massa de dados e documentos, vários tipos de sistemas foram concebidos, ao longo do tempo, para melhorar o controle do fluxo e extração de informações relevantes.

No que concerne ao armazenamento, Inmon (2008, 2016) sustenta que os *Data Warehouses* na atualidade têm tido problemas com dados não estruturados, visto que são tradicionalmente constituídos para suportar dados estruturados. Por outro lado, os *Data Lakes* – ou Lagos de Dados – têm se mostrado como uma tecnologia compatível com o volume, a velocidade e a diversidade dos dados não estruturados que circulam nas organizações. Uma vez processados nos *Data Lakes*, os dados podem então ser refinados e distribuídos para os *Data Ponds* (Lagoas de Dados) e após para os *Data Marts* (Repositórios Especializados ou Temáticos de Dados), percorrendo um circuito de tratamento de dados que têm se mostrado potencialmente interessante para a ecologia da informação organizacional. Entretanto, para Inmon (2008, 2015), a qualidade dos metadados e seu processamento semântico impede a plena decodificação ou tradução dos dados em informação relevante. Além disso, muitos dados podem ser perdidos ou duplicados pela precariedade do tratamento que estes têm recebido.

Ainda que uma nova geração de softwares para processamento e armazenamento de dados massivos e não estruturados tenham começado a empregar semântica ou ontologias para potencializar sua capacidade de recuperação de informação relevante no universo de dados, este uso encontra-se em um estágio inicial. O tratamento de dados pressupõe o uso de metadados, elementos que tem se mostrado essenciais para agregação de qualidade de modo a tornar os dados mais significativos. Para Baca (2008), os metadados possuem três recursos: conteúdo, contexto e estrutura. Para além da estrutura, o conteúdo e seu contexto exigem aspectos semânticos que possam ser interpretados e representados. Por sua vez, Inmon, O´Neil e Fryman sustentam que os metadados devem alinhar-se aos negócios, tanto no aspecto estratégico como operacional. Desta maneira, a geração de metadados deve levar em conta o contexto, condição essencial para consideração de aspectos semânticos. Entretanto, segundo o autor, "*infelizmente, a semântica dos sistemas de negócios tem sido amplamente ignorada*" (Inmon, O´Neil &

Fryman, 2008). Também para Eine, Jurisch e Quint, considerando as ontologias corporativas como um aspecto semântico, "*um desafio técnico é a aplicabilidade da correspondência de ontologia automatizada ao gerenciamento de Big Data. O atual estado da arte na correspondência de ontologias não suporta a construção de alinhamentos complexos entre ontologias a um grau satisfatório.*" (Eine, Jurisch & Quint, 2017). De acordo com estes autores, mesmo com uma expressiva capacidade computacional, os sistemas de interpretação e representação de informações não têm tido desempenhos satisfatórios neste sentido.

Muitos dados no ambiente corporativo consolidam-se em documentos que atendem a diferentes demandas e usuários. Também a velocidade e a quantidade de dados no ecossistema corporativo colocam problemas adicionais à consolidação de documentos estáticos. Além disso, determinados contextos – como departamentos e equipes internas – exigem dados e documentos consolidados de maneira customizada às suas necessidades, da mesma forma que, em outros contextos, tais documentos seriam customizados a outras equipes. Essa consolidação dinâmica e integrada a contextos diferentes têm sido o desafio de inúmeras organizações, sobretudo no que concerne ao aspecto semântico de tratamento de dados e metadados.

A importância da semântica no contexto corporativo pode ser constatada por Inmon, O´Neil e Fryman:

> A disciplina da semântica, portanto, é sobre tornar-se consciente dessas definições, suposições e natureza contextual dos dados. É também sobre tentar capturar esta informação para que os dados possam ser mais compreensíveis e também ser mais facilmente compartilhado, em toda a empresa e até mesmo externamente à empresa, quando for apropriado. (Inmon, O´Neil & Fryman, 2008)

Uma das representações de dados mais frequentes em sistemas empresariais se dá em RDF – *Resource Description Framework*, fato atestado por Inmon, O´Neil e Fryman, que o reconhece como um dos metadados semânticos de interesse para o ecossistema organizacional. Para Farid, Roatis e Ilyas, "o *RDF é de fato um formato comum para representar dados a partir de fontes de dados heterogêneas*." (Farid, Roatis & Ilyas, 2016). De acordo com estes últimos, softwares vêm surgindo de maneira a ajudar na extração de triplas de RDFs dos dados. Neste sentido, diversos outros formatos têm potencial para enriquecimento semântico de dados, sobretudo não estruturados, considerando os contextos de sua geração e uso.

1.1 Modelo conceitual de Soergel

Um modelo conceitual de representação de dados em documentos dinâmicos foi proposto por D. Soergel (2018), no qual dados extraídos de documentos são fragmentados

em unidades documentais e armazenados em bases de dados hipermídia. O processo de representação inclui a separação de dois tipos de elementos: estrutura e conteúdo. Segundo o autor, este modelo pode suportar as seguintes tarefas: Representar a estrutura do documento; Representar o conteúdo do documento; Fazer links com outros documentos; Armazenar muitos documentos numa mesma base de dados. Ademais, seu modelo propõe representar tipos de informações que servem a muitas funções, incluindo integração com workflow e anotações linguísticas para análise e aprendizado de máquina, dentre outras coisas. Em seu exemplo, foi sugerido uma decomposição de um relatório do *Banco Mundial*, como mostrado na figura 1:

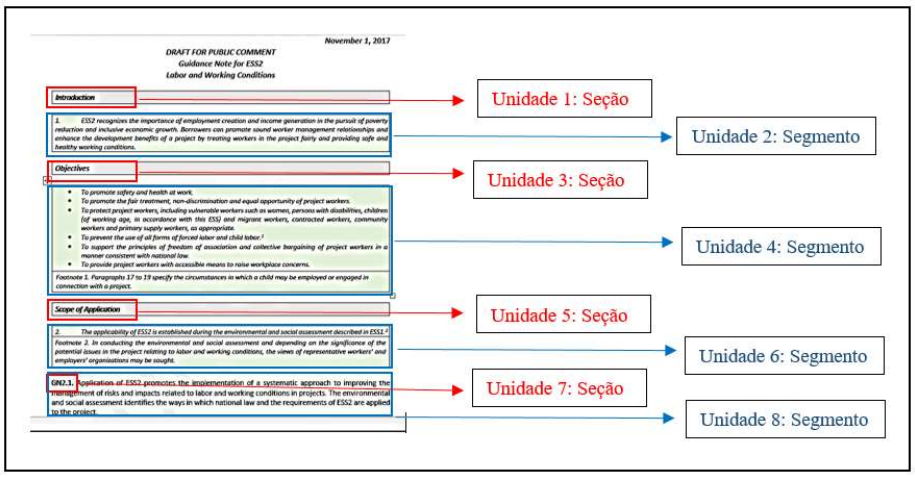

Figura 1: Aplicação do modelo conceitual de Soergel.

No exemplo acima, a estrutura é preservada decompondo-se em seções, que por sua vez possuem conteúdos que se decompõem em segmentos. Esta decomposição produz coleções de unidades a serem armazenadas em bases de dados hipermídia e, mediante consulta, os resultados retornados poderão diferir de acordo com as formulações de buscas. Desta maneira, na consulta, documentos dinâmicos variados poderão ser montados e exibidos ao usuário de modo a atender sua necessidade específica.

O exemplo apresentado no modelo de Soergel decompõe somente documento textual estruturado, no qual os blocos de dados no documento possuem demarcações espaciais específicas. Entretanto, o complexo ambiente corporativo possui dados diversos e não estruturados. Mais do que documentos textuais, também dados de multimídia compõem sua ecologia informacional. Soergel sustenta que seu modelo é genérico propõe-se a abrir muitas possibilidades, sendo possível testá-lo em contextos mais específicos ou complexos.

2 Objetivos

A pesquisa pretende estender o modelo de Soergel para aplicação em contexto corporativo típico, considerando dados não estruturados em formato não somente textual, mas também multimídia, como áudio e vídeo, dentre outros. Pretende também adicionar processos complementares ao modelo de Soergel, como destilação de dados e enriquecimento semântico de metadados, de modo a compor documentos dinâmicos com possibilidades de agregação a outros objetos, internos ou externos à organização. Serão disponibilizados exemplos de aplicação do modelo no ambiente organizacional.

3 Metodologia

Para o atingimento dos objetivos expostos, foi definido que a pesquisa será de caráter qualitativo e exploratório. A pesquisa também poderá ser caracterizada como *pesquisa-ação* (Mueller, 2007), na qual a pesquisa é orientada à busca de resultados específicos ou soluções inovadoras de problemas observados num ambiente empírico.

Após a revisão bibliográfica e a contextualização da pesquisa no ambiente corporativo, serão apresentados os seguintes recortes e passos nos quais o modelo será aplicado:

Ambientação

Definição dos casos de uso: serão apresentados casos típicos no ambiente corporativo que envolvam o uso de dados tal como demarcados no escopo da pesquisa.

Definição de tipos e fonte de dados

Serão definidos os tipos de dados no *Data Lake* a serem processados, a saber:

- Dados não estruturados
  - Multimídia – áudio vídeo e imagem
  - Dados textuais

Destilação dos dados

Será aplicado o processo de destilação de dados não estruturados, que consiste na decomposição dos dados, utilizando ferramentas de reconhecimento de padrões, comparações com a base de taxonomia e tesauro corporativo para identificação de termos, conceitos e demais aspectos. Além dos dados em si, também os metadados técnicos intrínsecos serão confrontados com metadados de negócios disponíveis na base de taxonomia corporativa. No processo de destilação serão utilizadas ferramentas de extração, como *Open Calais* (Thomson Reuters) e *AlchemyAPI* (IBM), bem como ontologias MPEG-7, dentre outras. Após a destilação, os dados serão alocados em *Data Ponds*.

Enriquecimento de metadados

Uma vez feita a decomposição por destilação nos *Data Ponds*, os metadados serão enriquecidos semanticamente com vocabulários ontológicos específicos, tal como RDF,

OWL, DC, ORE, dentre outros. O processo de enriquecimento semântico consistirá na inscrição de tais vocabulários nos metadados dos objetos e suas partes decompostas, obedecendo a critérios pré-determinados pelo modelo. Com isto, pretende-se permitir aos sistemas de informações corporativos a leitura dos metadados enriquecidos dos objetos, visando uma montagem de documentos dinâmicos, de modo a fazer referências a áreas, fontes, processos, atores e a outros objetos do ecossistema corporativo, além de enlaces a objetos na web, mediante processo de agregação. Estas possibilidades de agregações dos dados que comporão os documentos dinâmicos serão analisadas considerando os casos de uso expostos na pesquisa, onde será possível mensurar o desempenho dos processos de enriquecimento semântico e agregação de dados.

Armazenamento

Com a decomposição, destilação e enriquecimento semântico, pretende-se que os dados sejam, então, alocados em unidades nos *Data Ponds* e *Data Marts* corporativos, ficando disponíveis, assim, para consulta.

4 Resultados Esperados

A presente pesquisa, como já mencionado, pretende expandir o modelo proposto por Soergel a objetos não estruturados, como multimídia e outros, decompondo dados originais e submetendo-os a processos de destilação de dados e enriquecimento de metadados para agregação, de modo a permitir montagem de documentos dinâmicos. Como resultado, espera-se que o modelo estendido permita processar dados heterogêneos segundo os recortes estabelecidos e mediante os processamentos acima elencados, permitindo criação de valor na composição de documentos dinâmicos com agregações semânticas adicionados aos metadados.

Referências

Baca, M. (2008). *Introduction to Metadata*. (2nd. ed.) Los Angeles, CA: Getty Research Institute, CA.

Brocke, J. V., & Simons, A. (2014). *Enterprise Content Management in Information Systems Research*: foundations, methods and cases. Berlin: Springer-Verlag.

Eine, B., Jurisch, M., & Quint, W. (2017). Ontology-based big data management. *Systems*, v. 5 (3), 2017.

Farid. M., Roadis, A., & Ilyas, I. (2016). CLAMS: Bringing Quality to Data Lakes, *Proceedings of the 2016 International Conference on Management of Data*, June 26-July 01, 2016, San Francisco, CA.

Inmon, H. W., O´Neil, B., & Fryman. L. (2008). *Business Metadata*: capturing enterprise knowledge. Burlington, MA: Morgan Kaufmann Publishers.

Inmon, W. H., Imhoff, C., & Sousa, R. (2001). *Corporation information factory*. (2nd. ed.). New York, NY: Wiley & Sons.

Inmon, W. H., & Linstedt, D. (2015). *Data Architecture: a primer for the data scientist*. Waltham, MA: Elsevier.

Inmon, W. H. (2016). *Data Lake Architecture*: designing the data lake and avoiding the garbage dump. Basking Ridge, NJ: Technics Publications.

Marquesone, R. (2015). *Big Data*: técnicas e tecnologias para extração de valor nos dados. São Paulo: Casa dos Códigos Editora.

Mueller, S. M. (2007). *Métodos para a pesquisa em Ciência da Informação*. Brasília, DF: Thesaurus.

Rifkin, J. (2012). *A Terceira Revolução Industrial*: como o poder lateral está transformando a energia, a economia e o mundo. São Paulo: M. Books.

Schwab, K. (2016). *A Quarta Revolução Industrial*. São Paulo: Edipro.

Soergel, D. (2018). *Turning documents into active knowledge*: A universal document model with rich component indexing. III Seminário MHTX - Escola de Ciência da Informação. Universidade Federal de Minas Gerais – Escola de Ciência da Informação, Belo Horizonte, Brazil, 2018, 7-8 June.

Swanson, E. B. (2012). Information Systems. In: Bates M. (ed.). *Understanding information retrieval systems*: management, types, and standards. Boca Raton, FL: CRC Press.