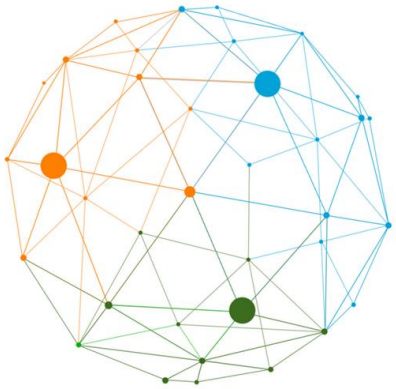


CDMWS

I COLÓQUIO DE DADOS, METADADOS E WEB SEMÂNTICA

14 A 15 DE DEZEMBRO DE 2017 | SÃO CARLOS

CONTEXTUALIZAÇÃO DE CONCEITOS TEÓRICOS NO PROCESSO DE COLETA DE DADOS DE REDES SOCIAIS ONLINE

CONTEXTUALIZING THEORETICAL CONCEPTS ON ONLINE SOCIAL NETWORKS DATA COLLECTING PROCESS

Fernando de Assis Rodrigues[1], Ricardo César Gonçalves Sant'Ana[2]

Resumo: O uso de serviço de redes sociais on-line suscitam preocupações na forma que informações dos indivíduos são compartilhadas, como, por exemplo, a partir do processo de coleta de dados de usuários que estão armazenados nas instituições proprietárias dos serviços. O objetivo deste estudo é estabelecer uma contextualização dos conceitos envolvidos no processo de coleta de dados disponibilizados por serviços de redes sociais online, a partir da análise de conteúdo realizada em documentos de cunho técnico-operacional e nos Termos de Uso e pela exploração das características das interfaces de coleta. Como metodologia, optou-se pela relação dos conceitos a partir da descrição do processo, com origem na análise de conteúdo dos documentos das redes sociais online para a delimitação das características e do funcionamento do processo de coleta de dados e, posteriormente, pela exploração das interfaces de coleta de dados, com intuito de delimitar elementos adicionais envolvidos com o contexto de coleta de dados. Como resultado se apresenta a descrição do processo de coleta de dados e a relação com conceitos de aportes interdisciplinares, relacionadas aos três ciclos de coleta propostos para sistematização da coleta e construção de modelos de dados. Concluiu-se que a coleta de dados é uma atividade com forte relação interdisciplinar e de cooperação, e envolve conceitos originários de diferentes áreas do

[1] Doutor e Mestre em Ciência da Informação pela Universidade Estadual Paulista (UNESP), pelo Programa de Pós-Graduação em Ciência da Informação, e-mail: fernando.orionx@gmail.com.

[2] Professor adjunto da Universidade Estadual Paulista - UNESP, Faculdade de Ciências e Engenharias - FCE, Campus de Tupã, em regime de dedicação exclusiva, onde é Coordenador Local do CENEPP-Centro de Estudos e Práticas Pedagógicas e Ouvidor Local. Professor do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista, Campus de Marília, e-mail: ricardosantana@marilia.unesp.br.

conhecimento, tornando-a complexa à compreensão de características processos de coleta de dados em sistemas de informação digitais – e espera-se que esta conceitualização inicial dos fundamentos seja subsídio suscitar a reflexão a novas investigações tanto de estudos do processo em si, mas também como uma orientação de base sobre estes temas.

Palavras-chave: Coleta de Dados. Dados. *Application Programming Interface*. Redes Sociais Online.

Abstract: The use of Online Social Network services raise concerns in the way information from individuals is shared, such as starting from the process of collecting data from users that are stored in the institutions that own this services. The purpose of this study is to establish a contextualization of the concepts involved in the data collection available at online social network services, based on the analysis of content in technical-operational documents and in Terms of Use, and by an exploration of the characteristics of the data collection interfaces. As a methodology, it was chose to start the relationship between concepts from the process description, originated on the content analysis of online social networks documents, to delimit the characteristics and operation of the data collection process and, after, the exploration of the data collection interfaces, in order to delimit additional elements involved with the context of data collection. As a result, its presented the description of data collection process and relationship with interdisciplinary concepts, related to the three collection cycles proposed for the systematization of data collection and construction of data models. It was concluded that data collection is an activity with a strong interdisciplinary and cooperative relationship, and involves concepts originating from different areas of knowledge, making it complex to the understanding of characteristic of data collection processes in digital information systems - and expects that this initial conceptualization will allow reflecting on further investigations of process itself and also as a basic guideline on these issues.

Keywords: Data Collecting. Data. Application Programming Interface. Online Social Networks.

1 INTRODUÇÃO

A formação e o desenvolvimento de redes de inter-relacionamento entre indivíduos de uma sociedade não é um fenômeno recente. Entrelaça-se com o desenvolvimento das sociedades e, nos países de influência Europeia, têm elementos da esfera pública da sociedade grega e da romana (HABERMAS, 1984) – onde ambas estimularam o desenvolvimento de ambientes para a exposição de ideias e discussões, tais como a Ágora e o Senado, e a formação de redes para o funcionamento administrativo do Estado (MALKIN, 2011).

As tecnologias relacionadas aos processos de coleta, de armazenamento e de recuperação de dados, associados à disponibilidade das redes de comunicação, tomam forma no uso de Tecnologias da Informação e Comunicação (TIC) e na ampliação da infraestrutura da Internet (FUMERO; VACAS; ROCA, 2007) – que possibilitaram o desenvolvimento de sistemas de informação digitais que suportam um conjunto cada vez maior de usuários conectados – o que é também parte de um fenômeno social e de uma sociedade entendida como uma “Sociedade em Rede” por Castells (2008), em que o principal ativo é a informação e as suas forças motrizes são as TIC.

É neste cenário que estão inseridas as redes sociais online, e seus sistemas de informação são desenvolvidos para suportar a coleta, o armazenamento e a recuperação de dados sobre os nós da rede e as suas ligações com outros pares (nós) e para associar estes sistemas à disponibilidade de serviços na Internet, com interfaces desenvolvidas para a troca de informações entre pares (BOYD; ELLISON, 2007).

O número de usuários destes sistemas de informação está em crescimento – com redes sociais online apresentando números superiores a um bilhão de usuários e milhares de agentes externos, conectados de forma concorrente (ao mesmo tempo) – e são novos *locus* para fenômenos de organização e reorganização social e cultural (BOYD, 2008, 2013; BOYD; ELLISON, 2007).

Entretanto, os serviços disponíveis pelas redes sociais online são, prioritariamente, propriedade de instituições privadas, e as trocas de informações entre usuários realizadas nestas redes (e o consequente compartilhamento de suas experiências) suscitam preocupações já existentes em outros contextos, tais como a exposição de conjuntos de dados sobre usuários para outras instituições, governos e, inclusive, para outros usuários; crimes sexuais e abusos contra a criança e a juventude; a perseguição online de pessoas (*cyberstalking*), e; ações e atividades resultantes de intolerância – conforme estudos de

Acquisiti e Gross (2006), Barnes (2006), Boyd (2008, 2013), Boyd e Ellison (2007), Dinev e Hart (2004), Fogel e Nehmad (2009), Krasnova et al. (2009), Tufekci (2007), Viseu et al. (2004), Young e Quan-Haase (2009) e Rodrigues e Sant'Ana (2015, 2016).

Neste sentido, as informações que permitem o acompanhamento sobre as características inerentes à coleta de dados por agentes externos estão descritas nos Termos de Uso e nos documentos em acervos de cunho técnico-operacional disponíveis nestes serviços. Este conjunto de documentos têm duplo papel para usuários de redes sociais online:

> […] pacificador, enquanto elemento que fortalece a percepção de segurança aos usuários, ao estabelecer os limites, as garantias legais sobre o que é realizado com conjuntos de dados pessoais por estes serviços de instituições privadas, e; como elemento de opacidade entre usuários e os serviços sobre o modus operandi do uso de dados pessoais, diluído em uma alta complexidade da rede e no volume e na variedade de ações e atividades passíveis de realização (Rodrigues e Sant'Ana, 2015, p. 245).

Neste contexto, o presente estudo ancora-se em uma intersecção teórica exploratória, de caráter qualitativo, com o objetivo de estabelecer uma contextualização dos conceitos envolvidos ao processo de coleta de dados disponibilizados por serviços de redes sociais online, a partir da análise de conteúdo sobre o processo de coleta de dados, realizada em documentos de cunho técnico-operacional e nos Termos de Uso, e pela exploração das características das interfaces de coleta.

O universo de pesquisa está delineado aos acervos e as interfaces de coleta de dados disponíveis nos serviços das redes sociais online, com amostragem delimitada as redes sociais online *Facebook*, *Twitter* e *LinkedIn*.

Os documentos e interfaces de coleta de dados analisadas estão vinculadas às versões:

- Para a rede social *Facebook*: a *Graph API*, nas versões 2.6 e 2.8;
- Para a rede social *Twitter*: a *REST API*, na versão 1.1;
- Para a rede social *LinkedIn*: a *REST API*, na versão 1.0.

2 METODOLOGIA

A metodologia adotada para descrever o processo de coleta de dados e para identificação dos conceitos foi a investigação por meio de (a) pesquisa exploratória, realizada por observação direta e não participante, a partir da identificação das características técnicas e operacionais das interfaces de coleta de dados, e (b) pesquisa bibliográfica, com análise de conteúdo dos documentos técnico-operacionais e dos Termos

de Uso para estabelecer a contextualização a partir da descrição das características da coleta de dados e estabelecer os vínculos com a literatura científica, com reflexões e considerações a partir de conceitos já desenvolvidos, como: de dados e ciclo de vida dos dados (SANT'ANA, 2016; SANTOS; SANT'ANA, 2015), de redes sociais e de redes sociais online (ADAMIC; ADAR, 2003; BOYD, 2008, 2013; BOYD; ELLISON, 2007; DURKHEIM, 1999; MORENO, 1953; RODRIGUES; SANT'ANA, 2015, 2016), de grafos sociais (BIGGS; LLOYD; WILSON, 1986; MORENO, 1953) e de modelos e estruturas para bancos de dados (CODD, 1981, 1990; DATE, 2016; INMON, 1996; KIMBALL; ROSS, 2011; PAPAZOGLOU, 2003; SILBERSCHATZ; KORTH; SUDARSHAN, 1999).

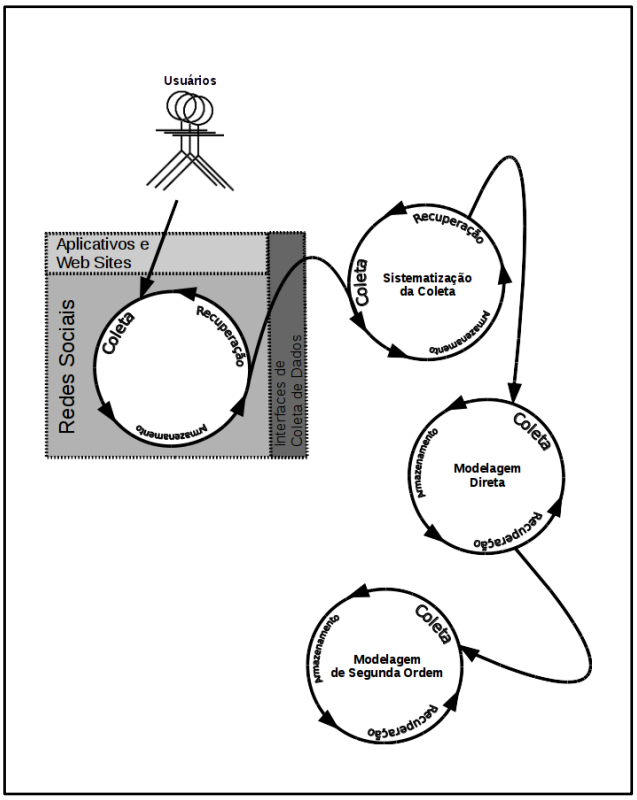
O processo de coleta adotado neste estudo está dividido em três ciclos, conforme Rodrigues (2017), onde o primeiro ciclo trata sobre a Sistematização da Coleta de dados de características identificadas do processo de coleta de dados de usuários de redes sociais online, como, por exemplo, a partir da sistematização de estruturas de coleta de dados sobre as características identificadas nos recursos tecnológicos e sobre dados gerados pelos procedimentos realizados por equipes de coleta.

O segundo ciclo está voltado à construção de uma camada de abstração para coleta, armazenamento e recuperação de conjuntos de dados já sistematizados no ciclo Sistematização da Coleta, denominado Modelagem Direta: uma modelagem de dados derivada das estruturas e da coleta de dados elaboradas no primeiro ciclo, concomitante aos conceitos de modelagem de dados para aplicação em Sistemas Gerenciadores de Banco de Dados (SGBD) (RODRIGUES, 2017).

Uma vez que na Modelagem Direta a aplicação deste modelo de dados está voltada realização de análises quantitativas, Rodrigues (2017) propõe um terceiro ciclo, denominado Modelagem de Segunda Ordem, que tem como objetivo a reorganização do modelo de dados adotado na Modelagem Direta para a geração de um modelo de dados, porém orientado às análises qualitativas, como uma forma de mitigar o processo de coleta, armazenamento e recuperação de dados sobre características do processo de coleta de dados de usuários de redes sociais online.

Os ciclos são interligados (Figura 1) e cada ciclo é composto por três etapas internas (coleta, armazenamento e recuperação), conforme o Ciclo de Vida dos Dados para Ciência da Informação (CVD) (SANT'ANA, 2016).

Figura 1: Caminho de elaboração dos ciclos.



Fonte: Autores.

Optou-se por relacionar conceitos com a coleta de dados a partir da descrição do processo, ou seja, a investigação tem origem na análise de conteúdo dos documentos das redes sociais online para a delimitação das características e do funcionamento do processo de coleta de dados e, na exploração das interfaces de coleta de dados disponíveis para delimitar demais elementos envolvidos com o contexto de coleta de dados (como estruturas, linguagens e tecnologias adotadas para a coleta).

O olhar a partir desta perspectiva foi delineado como uma tentativa de auxiliar investigadores a compreender conceitos envolvidos nos processos de coleta de dados em redes sociais online, pois são ambiente de cooperação interdisciplinar e de competências distintas aos participantes, tais como o desenvolvimento de atividades: a construção de modelos conceituais para armazenamento dos dados da coleta, a programação de algoritmos para interfaces de consulta a banco de dados, a sistematização do processo de coleta de dados e até em cenários específicos como, por exemplo, de identificação e de análise de potenciais ações e atividades de coleta de dados de usuários estabelecidas por agentes externos.

3 CONTEXTUALIZAÇÃO DOS CONCEITOS NO PROCESSO DE COLETA DE DADOS

As redes sociais online possuem um conjunto de serviços para auxiliarem os processos de comunicação e de inter-relacionamento de pessoas e de instituições participantes, elaboradas e mantidas por uma ou mais instituições, onde é disponibilizado acesso por meio de ferramentas em web sites ou por aplicativos e oferece acesso a parte dos dados armazenados de seus participantes para coleta por agentes externos.

Uma das características dos sistemas de informação das redes sociais online é a capacidade de identificação e distinção dos conjuntos de dados de cada participante da rede, incluindo dados relacionados aos seus atributos, as suas características e as suas relações com outros usuários e com demais conteúdos multimídia.

Os dados dos usuários disponíveis são formados pelo conjunto de dados e de metadados sobre o indivíduo, e sobre os seus relacionamentos com conteúdos e demais indivíduos. Possuem atributos que garantem a unicidade de cada indivíduo, denominados como identificadores, que são atributos referenciáveis dos indivíduos nos variados recursos multimídias disponíveis.

Portanto, um usuário sempre está relacionado com algum tipo de identificador: um documento de identidade na rede, de numeração única e intransferível – e ao identificar-se, a instituição proprietária da rede social online tem acesso aos conjuntos de dados com potencial de identificação deste usuário em outros domínios, como a data e hora de acesso; a geolocalização em um sistema de posicionamento global (como o *GPS*, acrônimo do termo inglês *Global Positioning System*); o número de *Internet Protocol (IP)*; informações sobre a rede de acesso; idiomas; entre outros.

Portanto, entende-se apropriado o uso do termo referenciado para descrever este tipo de usuário, que se relaciona com dados, tecnologias e sistemas de informação da rede social online, com potencial possibilidade de distinção de um referenciado dos demais, e a identificação de seus atributos em um ou mais domínios pelos códigos identificadores ou pelos dados do referenciado com potencial de identificação.

A forma de representação mais usual para visualizar relacionamentos sociais – o grafo social – tem origem nos estudos de sociometria, precursores no processo de explicitação das relações sociais entre indivíduos em visualizações. As regras do grafo social são baseadas na Teoria dos Grafos: estudo com origem na Matemática sobre conjuntos formados por objetos (vértices) e as relações dos objetos com o conjunto (arestas).

As formas geométricas circulares podem representar os referenciados em uma rede (denominados como vértices). O formato, o tamanho ou a coloração do vértice podem ser modificados na representação para, por exemplo, identificar padrões (como agrupamentos ou subgrupos) ou na aplicação de categorização, além de incluir o uso de rótulos para designar informações sobre cada vértice, como o nome ou o código identificador.

As ligações entre os vértices são explicitadas na forma de vetores retos, curvados ou angulados, denominadas como arestas. As arestas possuem início e término nas bordas de dois vértices distintos, onde cada ligação representa um vínculo entre dois referenciados e não há limites mínimos ou máximos estabelecidos para restringir a quantidade de vínculos entre referenciados.

A coleta de dados de referenciados diretamente nos bancos de dados destes serviços não é uma prática recorrente, pois podem ser potenciais canais passíveis de coletas de dados dos referenciados não autorizadas. É usual a estes serviços que os processos de coleta de dados ofereçam restrições de acesso e determinem quais conjuntos de dados estarão disponíveis. Portanto, oferecem aos agentes externos processos de coleta de dados baseados em uma camada de abstração mais alta (camada de visão): um banco de dados com uma seleção predefinida de dados dos referenciados.

Para a manutenção deste processo, é desenvolvido uma camada de abstração intermediária, baseada em restrições lógicas de acesso (camada lógica), com o intuito de criar regras de acesso aos conjuntos de dados dos referenciados.

Os dados estão armazenados em uma camada física, com o uso de softwares voltados à manipulação de dados, como os SGDBs. A camada física apresenta preocupações voltadas aos aspectos de armazenamento físico e manipulação dos dados.

O estabelecimento da infraestrutura da Internet como forma de distribuição de software e de transmissão de dados influenciou as características dos serviços das redes sociais online, como também a forma dos processos de coleta de dados nestes serviços.

São parte integrante de um conceito denominado como Software como Serviço, onde se destaca o maior enfoque nos benefícios dos serviços que são oferecidos pela adoção da tecnologia do que nos benefícios da aquisição de uma nova tecnologia *per si*, além de preocupações relativas à independência de plataformas tecnológicas específicas para o funcionamento do serviço.

Este enfoque no serviço e no uso da Internet influenciou parte dos serviços online, que passaram a apresentar interfaces de coleta de dados como um elemento intermediário de interoperabilidade de dados com agentes externos, sendo formado não só por algoritmos

e por serviços de interoperabilidade, mas também por acervo de documentos técnico-operacional e Termos de Uso, exemplos de requisições, ambiente para realização de testes e outros recursos informacionais para estabelecer as regras de acesso e a forma de operacionalização da coleta de dados.

Uma das formas encontradas de aplicação de interface de coleta de dados em serviços na Internet é por meio da adoção das Interfaces de Programação de Aplicativos (no singular, acrônimo *API*, do termo inglês *Application Interface Programing*):

> […] uma estrutura formal de regras e protocolos para proporcionar a interoperabilidade de conjunto de dados, por dois ou mais sistemas de informação, independentes de plataforma, de acesso público, privado ou misto, que utiliza padrões abertos ou fechados para o intercâmbio dos dados e contém documentação disponível na origem para o entendimento de todas as partes sobre o seu modo de operacionalização (RODRIGUES, 2017, p. 92).

Nas *APIs* dos serviços de redes sociais online, as requisições de coleta de dados de referenciados devem ter obrigatoriamente um ponto inicial de consulta como, por exemplo, o envio na solicitação com número identificador de um referenciado e, com isto, inicia-se o processo de requisição e de coleta dos conjuntos de dados da rede social online. Os referenciados são representados como pontos da rede (como os vértices) e os vínculos representam ligações entre dois referenciados e entre referenciados e conteúdos (como as arestas).

Este processo de coleta de dados ocorre pela aplicação de protocolos ligados ao contexto de *Wire Protocol*: no momento da transmissão de dados em uma rede, não é necessário que uma aplicação conheça os demais protocolos envolvidos na forma de transmissão de dados entre dois pontos, ou seja, não é necessário a compreensão de como os protocolos de transmissão de rede funcionam para realizar uma coleta de dados de uma *API*.

As *APIs* se beneficiam das estruturas e dos protocolos já existentes para a transmissão dos conjuntos de dados, como o Protocolo de Controle de Transmissão – Protocolo Internet (*Transmission Control Protocol – Internet Protocol – TCP/IP*).

No processo de coleta de dados de referenciados, o agente externo deve disparar uma requisição de acordo com as regras estabelecidas para a coleta de dados e com o formato do protocolo aceito, sendo comum o uso de sintaxes *REQUEST* originárias dos protocolos *Hyper Text Transfer Protocol (HTTP)* e *Hyper Text Transfer Protocol Secure (HTTPS)*.

A resposta da requisição é retornada ao agente externo pelas mesmas regras do protocolo utilizado em seu disparo e os conjuntos de dados retornados estarão codificados

em linguagem computacional, principalmente com uso de regras estabelecidas nas linguagens de marcação *eXtensible Markup Language (XML)* e *JavaScript Object Notation (JSON)*, e em alguns casos até em estruturas de arquivos com valores separados por vírgulas (acrônimo *CSV*, na língua inglesa *Comma-Separated Values*) e no uso de partes dos recursos disponíveis na linguagem *HyperText MarkUp Language (HTML)*.

As requisições ficam disponíveis aos agentes externos por meio de sistemas de entrada: serviços desenvolvidos para controlar o acesso aos conjuntos de dados, para controlar os diferentes perfis de acessos e níveis de permissões, e para o acesso a conjuntos de dados de outras temáticas, como coletas de dados de análises estatísticas e de dados de variáveis da configuração das contas dos referenciados.

As interfaces de coleta de dados e das *APIs* variam entre si – pois cada rede social online tem seu acervo documental, Termos de Uso, requisições e conjuntos de dados disponíveis e organizados de forma distinta – e variam no tempo – pois cada rede social online pode oferecer e pode remover funcionalidades da *API* em cada versão.

Nas etapas de investigação do processo de coleta de dados das redes sociais online, é importante ao pesquisador sistematizar um modelo de dados para formalizar e estruturar os conjuntos de dados coletados sobre as características de cada *API*, dos sistemas de entrada adotados, das autorizações e das permissões de acesso, das requisições, dos protocolos, da forma de recuperação e das linguagens utilizadas para explicitar os conjuntos de dados.

O conhecimento e habilidades em (a) aplicativos e ferramentas que permitam sistematizar modelos de dados sobre a coleta na forma da tríade Tabela/Coluna/Linha e (b) em instrumentos para a elaboração de formulários e outros mecanismos que facilitem o processo de coleta são elementos importantes para a construção de um panorama com as características gerais dos processos de coleta de dados, e quais são as características inerentes a cada processo de coleta de dados que as diferem das demais.

Existem diferentes abordagens para a elaboração de modelos de dados, em que cada abordagem possui um conjunto de técnicas e de conceitos para a construção de abstrações. Cada abordagem propõe formas e regras próprias sobre como elaborar modelos e estruturas de organização e aplicá-las aos dados. Também contam com instrumentos, próprios ou desenvolvidos por terceiros, para manipular a estrutura e os dados.

Na coleta de dados, destaca-se o uso do Modelo Entidade-Relacionamento (MER): formas e regras para estruturar modelos e aplicá-los aos dados coletados, onde utilizam-

se objetos para explicitar as estruturas de conjuntos de dados, denominados como entidades. As entidades têm outros elementos disponíveis que permitem relacioná-las entre si.

No MER, cada entidade é formada por uma tabela, incluindo a delimitação de um conjunto de atributos para a entidade (colunas), desenhado para representar suas características.

Por exemplo, no modelo de dados desenvolvido para coleta de dados de uma *API* de rede social online, os dados dos referenciados podem ser relacionados uma tabela nomeada ‘referenciado’, e as colunas da tabela ‘referenciado’ são formadas pelos atributos da entidade, como, por exemplo o ‘identificador’, o ‘nome’, o ‘e-mail’, entre outros. Na fase de armazenamento de dados, os dados de cada referenciado irão compor uma nova linha na tabela ‘referenciado’ e cada atributo será armazenado em uma coluna.

Portanto, para os elementos da tríade ‘Tabela, Coluna e Linha’: a Tabela representa um objeto do mundo real, com características que o distingue de outros objetos, as Colunas são as características intrínsecas do objeto, a Linha representa dos conjuntos de dados de um único objeto, contendo valores para cada coluna de uma entidade específica.

Cada coluna pode ter pré-determinada a aceitação ou não a um tipo de valor, possuindo restrições individuais dos tipos de valores que poderão ser armazenados na coluna, que podem ser: simples, com seu valor sendo formado por elementos básicos, como uma data, um número ou um texto, ou compostas, com seu valor formado por um subconjunto de colunas (colunas que contém duas ou mais colunas).

No MER, as restrições de valores para as colunas são denominadas tipos de dado, sendo comum aos SGBDs o uso dos tipos de dados:

- *boolean*, para armazenamento de números binários (como as condições sim ou não, verdadeiro ou falso, ligado ou desligado, entre outras);
- *integer*, para armazenamento de números inteiros;
- *float*, para armazenamento de números racionais finitos;
- *currency*, para armazenamento em formato de moeda;
- *string*, para armazenamento de símbolos, com tamanho reservável de forma fixa ou variável, e;
- *binary object*, para armazenamento de símbolos com extensão maior que o limite do tipo de dado *string*.

Algumas dessas nomenclaturas variam de acordo com especificações de cada instituição desenvolvedora de SGBD, mas com restrições equivalentes.

Uma ou mais colunas podem ser marcadas com identificadores especiais para garantir a unicidade e a identificação de cada linha de uma tabela: as chaves identificadoras.

As chaves identificadoras podem ser: (a) superchaves, que identificam unicamente uma linha (como o e-mail do referenciado em uma rede social online), (b) chaves candidatas, quando uma coluna possui potencial de identificação da linha.

Na elaboração do modelo de dados, superchaves e chaves candidatas podem ser escolhidas como chaves primárias: formam as colunas para a identificação única de cada linha na tabela, evitando a duplicidade ou redundância de linhas. As chaves primárias também podem ser formadas por um identificador único artificial, como, por exemplo, um número sequencial único para identificar cada linha.

Os identificadores podem se relacionar com demais linhas, em diferentes tabelas. Neste sentido, são denominados chaves estrangeiras: são chaves primárias que são inseridas como colunas em outras tabelas, mas qualificadas como chave estrangeira. Quando os valores da chave primária e estrangeira são iguais, e este procedimento formaliza o relacionamento entre as linhas das tabelas.

Nas redes sociais online, observa-se o uso das chaves primárias e estrangeiras para estabelecer o relacionamento entre diferentes entidades, tais como: o identificador do referenciado, que é uma chave primária formada por um identificador único artificial, e que seu valor aparece como chave estrangeira quando relacionado a outras entidades, como uma fotografia, uma mensagem, um vídeo, uma relação com um referenciado, entre outras.

A forma do relacionamento pode variar de acordo com a cardinalidade, ou seja, a explicitação lógica para determinar a quantidade de linhas que serão relacionadas entre as tabelas pelas chaves primária e estrangeira.

A cardinalidade é representada por um conjunto fixo de símbolos, sendo:

- ‘1-para-1’, quando o relacionamento entre as linhas das tabelas é permitido, no máximo, entre duas linhas (uma de cada tabela);
- ‘1-para-N’, quando a linha da tabela de origem pode relacionar-se com uma ou mais linhas da tabela de destino, e;
- ‘N-para-N’, quando linhas das tabelas de origem e destino podem se relacionar diversas vezes.

O MER possui instrumentos para a representação de tabelas, colunas e qualificadores, tipos de dado, identificadores, relacionamentos e cardinalidade de seus elementos. Destacam-se como instrumentos principais a elaboração de Diagrama

Entidade-Relacionamento (DER) e de Dicionário de Dados (DD).

O DER pode expressar elementos de um modelo de dados assim como os grafos sociais: a forma geométrica de retângulo é configurada para representar as tabelas (entidades), com cabeçalho contendo o nome da tabela e, na sequência, a lista com os nomes das colunas, os tipos de dado, as restrições de conteúdo, os símbolos especiais para diferenciar chaves primárias, chaves estrangeiras e colunas de preenchimento obrigatório, as relações e a cardinalidade.

O Dicionário de Dados (DD) é um instrumento elaborado para descrever as características das tabelas e das colunas que serão criadas pelo modelo de dados, incluindo um detalhamento sobre os tipos de dado, o tamanho máximo de armazenamento para cada coluna, a obrigatoriedade de inserção de valores na criação de novas linhas, as restrições de chave primária e a descrição sobre o conteúdo esperado no valor da coluna.

O DER e o DD são elementos importantes para o desenvolvimento do modelo de dados, no planejamento e na manutenção de suas características.

Para aplicar o modelo de dados em um SGBD é necessário adotar uma linguagem para explicitar as estruturas das tabelas, colunas e relacionamentos e, posteriormente, manipular os dados que serão coletados.

O uso da linguagem *Structured Query Language (SQL)* é proeminente no contexto de aplicação do MER, principalmente por ser uma linguagem disponível nos principais SGBDs e conter instruções concomitantes às regras de elaboração de um modelo de dados baseado em entidades e relacionamentos.

Portanto, é importante para as equipes envolvidas na coleta de dados conhecer noções da linguagem *SQL*, como o seu uso para explicitação do modelo de dados elaborado na forma de um MER, incluindo códigos-fonte desenvolvidos nesta linguagem para estruturas de armazenamento da coleta de dados, bem como conhecer funcionalidades de consulta, como as aplicações para visualização de conjuntos de dados armazenados – parte das suítes de ferramentas de manipulação de SGBDs e fundamentais para a realização de testes da estrutura do modelo de dados e do processo de coleta.

Também são adotadas linguagens de programação para a construção de algoritmos automatizados de coleta de dados e para realização de testes como: a delimitação das etapas do processo automatizado de coleta de dados, das requisições e do posterior armazenamento em SGBDs.

No caso de um estresse de informações, por fatores como a quantidade de dados armazenados a partir da execução da coleta ou pela necessidade de parte do modelo de

dados não poder ser acessível a determinadas audiências, os SGDBs possuem elementos para auxiliar o processo de segmentação destes conjuntos de entidades acessíveis, denominadas visões. As visões são formadas por estruturas pré-selecionadas pelos responsáveis do banco de dados, que podem incluir a elaboração de DER e de DD próprios.

As visões são compostas por uma pré-seleção de tabelas, de colunas, de linhas e de relacionamentos do modelo de dados, com capacidade de proteção no acesso às entidades do nível lógico e físico. O acesso e a coleta dos dados contidos nas visões são idênticas ao acesso e a coleta de dados diretamente em tabelas, exceto que, nas visões não é possível inserir, alterar ou excluir linhas.

Outra forma de se apropriar das visões é no seu uso para elaborar uma camada de abstração mais alta de acesso aos conjuntos de dados disponíveis, reorganizados com a finalidade de exibir, aos usuários, novas percepções a partir dos dados armazenados, como, por exemplo, construções de múltiplas visões de um mesmo conjunto de tabelas, voltadas a atender demandas específicas, como na formação de um *Data Warehouse*.

O *Data Warehouse (DW)* é a aplicação de um modelo de dados em um SGBD, orientado a assuntos, variável no tempo e não-volátil para auxiliar ao processo de tomada de decisão. A sua estrutura no modelo de dados difere-se das demais pois não é voltada ao uso e apoio de atividades diárias de uma instituição (transacionais) e, portanto, é estruturado com enfoque específico na realização de consultas orientadas por uma demanda de análise específica (analítica), organizada por dimensões.

Sua forma é dimensional e contém dois elementos de base: o fato e a dimensão. Além disso, é formado por um conjunto de tabelas, representando os fatos e as dimensões, onde cada conjunto interligado de fatos e dimensões é um *Data Mart.*

O *Data Mart (DM)* é um conjunto de dados flexível, reorganizados a partir de estruturas de um banco de dados transacional existente, apresentado em um modelo de dados dimensional – mais adaptável a realização de consultas do que em modelos de dados voltados à operacionalização de atividades transacionais – em que cada *DM* representa conjuntos de dados de um processo e suas dimensões de análise como, por exemplo, a mensagem e suas dimensões referenciado, data, horário e local.

A elaboração de um modelo de dados de um *DM* é dividida em duas fases: (a) a identificação de colunas que sejam fatos: primária, central, e é ponto de partida na construção de um modelo dimensional, onde serão armazenadas as colunas quantificáveis e as chaves estrangeiras das tabelas de dimensões; e (b) a elaboração das dimensões: componentes essenciais, contendo colunas com descrições textuais para cada uma das

chaves estrangeiras da tabela fato.

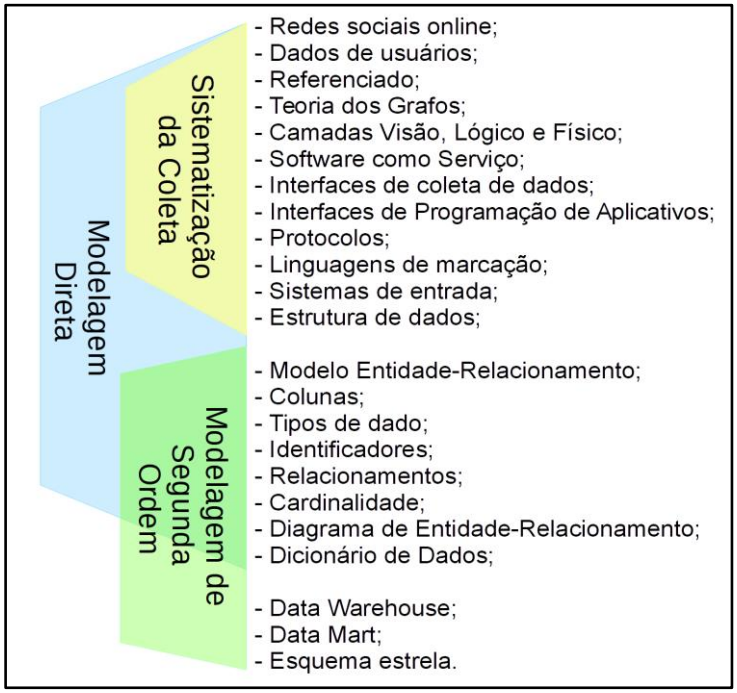
As dimensões serão parte integrante de subsídios para a construção de interfaces de consultas aos conjuntos de dados do *DM* e as colunas das tabelas de dimensão servirão como recursos de localização e de refinamento na recuperação dos dados.

O modelo de representação de *DM* que contém a tabela de fato circundada pelas tabelas de dimensões é conhecido como esquema estrela (no idioma inglês, *star scheme*).

Para esta camada de abstração, também é importante às equipes envolvidas na coleta de dados o conhecimento de noções da linguagem *SQL* e de linguagens de programação – como o uso destas linguagens para a explicitação e a construção de modelos de dados dimensionais de *DW* em um SGBD, incluindo a elaboração de códigos-fonte nesta linguagem para reorganizar dados coletados em *DMs* e o uso de aplicativos e algoritmos de testes como para a validação do processo de reorganização dos conjuntos de dados já armazenados.

A Figura 2 apresenta uma síntese dos conceitos envolvidos ao processo de coleta de dados em interfaces de dados disponíveis nos serviços de redes sociais online.

Figura 2: Relacionamento entre ciclos de coleta de dados e conceitos



Fonte: Autores.

Cada ciclo demanda um conjunto de conceitos próprios ao contexto de suas atividades, e no ciclo de Sistematização da Coleta estão relacionados conceitos sobre os temas: Redes sociais online, Dados de usuários, Referenciado, Teoria dos Grafos, Camadas Visão, Lógico e Físico, Software como Serviço, Interfaces de coleta de dados,

Interfaces de Programação de Aplicativos, Protocolos, Linguagens de marcação, Sistemas de entrada e Estrutura de dados.

No ciclo de Modelagem Direta, por desempenhar papel intermediário, são relacionados os temas do ciclo de Sistematização da Coleta e mais oito temas: Modelo Entidade-Relacionamento, Colunas, Tipos de dado, Identificadores, Relacionamentos, Cardinalidade, Diagrama Entidade-Relacionamento e Dicionário de Dados.

No ciclo de Modelagem de Segunda Ordem são congregados onze temas: Modelo Entidade-Relacionamento, Colunas, Tipos de dado, Identificadores, Relacionamentos, Cardinalidade, Diagrama Entidade-Relacionamento, Dicionário de Dados, *Data Warehouse*, *Data Mart* e Esquema Estrela.

4 CONSIDERAÇÕES FINAIS

A coleta de dados envolve um conjunto de ações e atividades estabelecidas em um projeto ou processo, desenvolvidas e aplicadas por humanos, máquinas ou por ambos, com tempo de duração e ciclos pré-definidos, auxiliado por métodos, protocolos, tecnologias e linguagens disponíveis, no intuito de obter de forma sistemática dados disponíveis na recuperação de uma ou mais fontes – incluindo o arranjo dos dados coletados por metadados para organização e representação e a aplicação de modelos de dados para estruturar os resultados da coleta à posterior fase de armazenamento.

As etapas da coleta de dados têm forte caráter e necessidade de cooperação interdisciplinar, o que reflete nos aportes de conceitos originários de diferentes áreas do conhecimento – como da Ciência da Computação, da Matemática e da Sociologia – e na formação de equipes interdisciplinares para processos de coleta de dados, principalmente derivada da necessidade de conhecimentos prévios de teorias, de ferramentas e de técnicas destas diferentes áreas do conhecimento.

Espera-se que esta conceitualização inicial dos fundamentos relacionados ao processo de coleta de dados de redes sociais online seja subsídio suscitar a reflexão a novas investigações que envolvem não só estudos do processo em si, mas também seja uma orientação de base para temas e conceitos na realização da coleta de dados em sistemas de informação digitais.

REFERÊNCIAS

ACQUISTI, A.; GROSS, R. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In: DANEZIS, G.; GOLLE, P. (Eds.). **Privacy Enhancing Technologies**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. v. 4258p. 36–58.
ADAMIC, L. A.; ADAR, E. Friends and neighbors on the Web. **Social Networks**, v. 25, n. 3, p. 211–230, jul. 2003.
BARNES, S. B. A privacy paradox: Social networking in the United States. **First Monday**, v. 11, n. 9, 4 set. 2006.
BIGGS, N.; LLOYD, E. K.; WILSON, R. J. **Graph theory, 1736-1936**. Oxford [Oxfordshire] ; New York: Clarendon Press, 1986.
BOYD, D. Facebook's Privacy Trainwreck: Exposure, Invasion, and Social Convergence. **Convergence: The International Journal of Research into New Media Technologies**, v. 14, n. 1, p. 13–20, 1 fev. 2008.
BOYD, D. **Making sense of teen life: Strategies for capturing ethnographic data in a networked era**. 2013.
BOYD, D. M.; ELLISON, N. B. Social Network Sites: Definition, History, and Scholarship. **Journal of Computer-Mediated Communication**, v. 13, n. 1, p. 210–230, out. 2007.
CASTELLS, M. et al. **A sociedade em rede**. São Paulo: Paz e Terra, 2008.
CODD, E. F. Data models in database management. **ACM Sigmod Record**, v. 11, n. 2, p. 112–114, 1981.
CODD, E. F. **The relational model for database management: version 2**. Reading, Mass: Addison-Wesley, 1990.
DATE, C. J. **The new relational database dictionary: a comprehensive glossary of concepts arising in connection with the relational model of data, with definitions and illustrative examples: [terms, concepts, and examples]**. Sebastopol, CA: O´Reilly, 2016.
DINEV, T.; HART, P. Internet privacy concerns and their antecedents – measurement validity and a regression model. **Behaviour & Information Technology**, v. 23, n. 6, p. 413–422, nov. 2004.
DURKHEIM, E. **Da divisão do trabalho social**. Tradução Eduardo Brandão. 2. ed. São Paulo: Martins Fontes, 1999.
FOGEL, J.; NEHMAD, E. Internet social network communities: Risk taking, trust, and privacy concerns. **Computers in Human Behavior**, v. 25, n. 1, p. 153–160, jan. 2009.
FUMERO, A.; VACAS, F. S.; ROCA, G. **Web 2.0**. [s.l.] Fundación Orange, 2007.
HABERMAS, J. **Mudança estrutural da Esfera Pública**. 1. ed. Rio de Janeiro, Brasil: Tempo Brasileiro, 1984.
INMON, W. H. **Building the data warehouse**. 2. ed. New York: Wiley Computer Pub, 1996.
KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit The Complete Guide to Dimensional Modeling**. New York, Estados Unidos da América: John Wiley & Sons, 2011.
KRASNOVA, H. et al. Privacy concerns and identity in online social networks. **Identity in the Information Society**, v. 2, n. 1, p. 39–63, 1 dez. 2009.
MALKIN, I. (ED.). **Greek and Roman networks in the Mediterranean**. London: Routledge, 2011.
MORENO, J. L. **Who shall survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama**. New York: Beacon House Inc., 1953. v. 58
PAPAZOGLOU, M. P. **Service-oriented computing: Concepts, characteristics and directions**. Proceedings of the Fourth International Conference on Web Information

Systems Engineering. **Anais**... In: FOURTH INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS ENGINEERING (WISE'03). IEEE, 2003.
RODRIGUES, F. DE A. **Coleta de dados em redes sociais: privacidade de dados pessoais no acesso via Application Programming Interface**. Tese—Marília, Brasil: Universidade Estadual Paulista, 3 mar. 2017.
RODRIGUES, F. DE A.; SANT'ANA, R. C. G. Uso de taxonomia sobre privacidade para identificação de atividades encontradas em termos de uso de redes sociais. Actas del XII Congreso ISKO España y II Congreso ISKO España y Portugal. **Anais...**: 12. In: XII CONGRESSO ISKO ESPAÑA E II CONGRESSO ISKO ESPAÑA-PORTUGAL. Murcia, Espanha: International Society for Knowledge Organization, 19 nov. 2015.
RODRIGUES, F. DE A.; SANT'ANA, R. C. G. Use of Taxonomy of Privacy to Identify Activities Found in Social Network's Terms of Use. **Knowledge Organization**, v. 43, n. 4, p. 285–295, 2016.
SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, v. 21, n. 2, p. 116, 20 dez. 2016.
SANTOS, P. L. V. A. DA C.; SANT'ANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, v. 42, n. 2, p. 199–209, 27 jan. 2015.
SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Database system concepts**. 6. ed. New York: McGraw-Hill, 1999.
TUFEKCI, Z. Can You See Me Now? Audience and Disclosure Regulation in Online Social Network Sites. **Bulletin of Science, Technology & Society**, v. 28, n. 1, p. 20–36, 27 dez. 2007.
VISEU, A.; CLEMENT, A.; ASPINALL, J. Situating privacy online: Complex perception and everyday practices. **Information, Communication & Society**, p. 92–114, 2004.
YOUNG, A. L.; QUAN-HAASE, A. **Information revelation and internet privacy concerns on social network sites: a case study of facebook**. Proceedings of the fourth international conference on Communities and technologies. **Anais**... In: 4TH INTERNATIONAL CONFERENCE ON COMMUNITIES AND TECHNOLOGIES. University Park, Estados Unidos da América: ACM Press, 2009.