DOI: 10.5433/1981-8920.2016v21n2p116

CICLO DE VIDA DOS DADOS: UMA PERSPECTIVA A PARTIR DA CIÊNCIA DA INFORMAÇÃO

CICLO DE VIDA DEL DATOS: UNA PERSPECTIVA DESDE LA CIENCIA DE LA INFORMACIÓN

Ricardo César Gonçalves Sant'Ana*

RESUMO

Introdução: O acesso e uso dos dados como fator chave de sucesso tem se estendido as mais diversas áreas do saber e do fazer da sociedade hodierna. Faz-se necessário o desenvolvimento de uma perspectiva que apresente fases e fatores envolvidos nestes processos, fornecendo uma estrutura inicial de análise que permita a organização de esforços, competências e ações relacionadas ao ciclo de vida dos dados.

Objetivo: Este artigo parte de uma proposta de um novo olhar para o Ciclo de Vida dos Dados, que pressupõe, como elemento central, os próprios dados, amparando-se nos conceitos e contribuições que a Ciência da Informação pode proporcionar, sem abrir mão da reflexão sobre o papel de outras áreas chave como a Ciência da Computação.

Metodologia: Os procedimentos metodológicos consistiram em pesquisa bibliográfica e análise de conteúdo para descrever as fases e fatores relacionados ao Ciclo de Vida dos Dados, tecendo reflexões e considerações a partir de contexto já consolidado no desenvolvimento de sistemas que possam corroborar com a ideia de centralidade dos dados.

Resultados: Como resultados apresentam-se as fases de coleta, armazenamento, recuperação e descarte, permeadas por fatores transversais e presentes em todas as fases: privacidade, integração, qualidade, direito autoral, disseminação e preservação, compondo um Ciclo de Vida dos Dados.

Conclusões: O contexto atual de disponibilidade de grandes volumes de dados, com grande variedade e em velocidades que propiciam o acesso em tempo real, configurando o assim denominado *Big Data* requer novos olhares para os processos de acesso e uso de dados. A Ciência da Informação pode oferecer um novo enfoque, agora centrado nos dados, e contribuir para a otimização do Ciclo de Vida dos Dados como um todo, ampliando as pontes entre os usuários e os dados que necessitam.

Palavras-chave: Ciclo de Vida dos Dados. Ciência da Informação. *Big Data*.

*Doutor em Ciência da Informação pela Universidade Estadual Paulista. Docente do Programa de Pós-Graduação em Ciência da Informação – FFC – UNESP Marília – SP. E-mail: ricardosantana@tupa.unesp.br

1 INTRODUÇÃO

O acesso a dados vem transformando todas as áreas de atuação humana, com especial crescimento nos últimos anos em função do aumento exponencial de alternativas para coleta, armazenamento e recuperação de dados, superando, inclusive, nossa capacidade para lidar com estas novas perspectivas de volume, variedade e velocidade de acesso a dados, cenário que vem sendo definido como *Big Data*.

> Big data se refere a dados que são grandes demais para um único servidor, muito diversos para se adequar a uma base de dados estruturada em linhas e colunas, ou cujo fluxo seja tão intenso que não permita adequação a um data warehouse estático (DAVENPORT, 2014).

Este cenário além de indicar um novo conjunto de possibilidades de uso e consequências desta alta disponibilidade, ainda que passiva, de dados, aponta para novos requisitos na construção de pontes entre os usuários, a partir de suas necessidades informacionais, e este contexto de potencialidades, para que se possa reduzir eventuais efeitos colaterais como a assimetria informacional (AKERLOF,1970) que pode ser tão danosa para os indivíduos e para a sociedade como um todo, quanto os ganhos que o acesso massivo a dados pode trazer.

Esta potencial assimetria se sustenta na necessidade de camadas de recursos tecnológicos para uso destes dados que tendem a ser mais profundas e complexas com o aumento das características de *Big Data*, o que justifica um olhar para as entranhas desta tecnologia. Neste cenário emerge a necessidade de uma estrutura básica em que se possa, ao menos, contextualizar momentos, características e requisitos bem como fatores que permeiam estes momentos. É necessário pensar em uma estrutura que permita a percepção destes momentos e características, que no caso dos dados podem ser considerados como cíclicos, proporcionando a reflexão sobre um ciclo de vida dos dados.

Ao se resgatar aspectos relacionados a uma perspectiva histórica do desenvolvimento de soluções de Tecnologia da Informação, percebe-se que

com a difusão de plataformas digitais, amplia-se a necessidade de padronização para modelagem e demais fases do projeto, com o surgimento de diversas abordagens, das quais pode se destacar o Projeto Estruturado (STEVENS; MYERS; CONSTANTINE, 1974; YOURDON; CONSTANTINE, 1978), a Análise Estruturada (DEMARCO, 1979), Análise Estruturada Moderna (YOURDON, 1989), Análise Essencial (MCMENAMIN; PALMER, 1984) ou mesmo Análise Orientada a Objetos - AOO (COAD; YOURDON, 1990), manteve-se sempre uma distinção entre as definições sobre os dados e as definições sobre as funcionalidades.

Esta distinção justifica-se pelo fato de que as funcionalidades apresentam uma dinamicidade em suas definições muito maior que as definições sobre os dados, já que estas estão mais próximas das camadas de relação com os usuários bem como com as camadas de integração com o suporte e suas características. Mesmo na AOO que propunha uma ruptura nesta dissociação, manteve em seu elemento de estruturação, a classe, que por sua vez definia como os objetos seriam instanciados, uma divisão na modelagem em que se definiam em uma seção os atributos e em outra os métodos, ou comportamento.

Com o contexto do *Big Data* se consolidando como uma realidade, ganha destaque, portanto, a necessidade de se ter um enfoque especial na questão dos dados e, assim, propõe-se, que se considere como ponto de referência os dados e não as demais dimensões conforme outros ciclos de vida dos dados (DATA DOCUMENTATION INITIATIVE, 2004; UNIVERSITY COLLEGE LONDON, 2012; HUMPREY, 2006; PENNOCK, 2007; FERDERER, 2001; INTERAGENCY WORKING GROUP ON DIGITAL DATA, 2009; VANDERBILT UNIVERSITY MEDICAL CENTER, 2005; DATA OBSERVATION NETWORK FOR EARTH, 2013; MATERIAL DATA MANAGEMENT CONSORTIUM, 2013; DIGITAL CURATION CENTER, 2013; CHEN; CHEN; LIN, 2003; SANT'ANA, 2013). Portanto, ao se considerar este modelo de ciclo de vida deve-se ter em mente que tanto os momentos quanto os fatores envolvidos tomando como elemento central os dados.

2 CICLO DE VIDA DOS DADOS

A ciência da Informação pode e deve contribuir para que este cenário de acesso e uso intenso de dados se desenvolva da melhor maneira possível, buscando identificar e estudar fatores e características que propiciem ampliação do equilíbrio entre os atores envolvidos no processo e a máxima otimização do uso dos dados.

Nesta tarefa se faz necessário estruturar esta análise, e para tanto, propõe-se a utilização de uma delimitação de fases (momentos em que distintas necessidades e competências são necessárias) envolvidas no acesso e uso dos dados, mantendo-se como ponto central os próprios dados e para tanto se propõe o uso do Ciclo de Vida dos Dados - CVD como forma de evidenciar os diferentes momentos e fatores envolvidos neste processo.

Em um primeiro momento é preciso obter os dados que podem ser utilizados para atender uma necessidade específica ou uma demanda prevista de informações sobre um determinado contexto. Neste primeiro momento são necessários esforços para que se possa estabelecer um plano de ação, análise da viabilidade bem como a execução da **coleta** dos dados. Entre outras questões fundamentais desta fase pode-se destacar: Qual é o escopo da necessidade informacional? Que tipo de resultado se espera? Com quais características? Quais são os dados necessários? Onde estão as fontes para estes dados? Como os dados podem ser coletados? Em que formato estão? Quais são os tratamentos necessários para que fiquem adequados ao que se precisa? A coleta destes dados não proporciona risco de privacidade para os indivíduos ou entidades referenciados por eles? Elementos, que em alguns casos poderiam ser considerados como secundários, que permitam a integração entre os diversos dados coletados estão sendo obtidos? Como avaliar sua integridade física e lógica, além de outros elementos que garantam sua qualidade? Como identificar sua procedência? Têm-se o direito ou permissão de coletar estes dados? Estão sendo coletados dados que permitam que estes venham a ser identificáveis e recuperáveis em um momento futuro? Estão sendo coletados dados que propiciem a manutenção e acesso a eles no futuro caso venham a ser armazenados?

Assim, percebe-se uma fase em que são necessárias competências específicas, ainda que não totalmente dependentes de um conhecimento profundo sobre as tecnologias digitais, mas muito próximo da necessidade informacional que motiva a coleta. Portanto, nesta fase, tanto o usuário, quanto àqueles que detêm conhecimentos advindos da Ciência da Informação quanto da Ciência da Computação são personagens importantes e trabalhando em conjunto podem tornar o processo mais eficiente.

Uma vez obtidos os dados, estes podem ser utilizados para um fim imediato e descartados, o que, como veremos, pode ser considerado como outra fase. No entanto, pode ser necessário e útil manter estes dados disponíveis de alguma forma para acesso futuro.

Com a evolução dos recursos digitais, o custo de aquisição e manutenção de suportes digitais é cada vez mais acessível e viabiliza que a decisão por manter os dados seja cada vez mais fácil e desejada. Neste momento passa-se então a uma segunda fase que é aquela em que os esforços são no sentido de manter estes conteúdos em um determinado suporte. Na computação este processo denominado de persistência dos dados leva a uma série de preocupações e aspectos que devem ser detalhadamente planejados. Tem inicio então uma fase em que o objetivo é **armazenar** estes dados. Dentre outras questões fundamentais desta fase pode-se destacar: Quais são os dados disponíveis? Quais destes dados serão armazenados? Qual estrutura (física e lógica) será utilizada para seu armazenamento? Como garantir a permanência dos dados complementares sobre a coleta para que se tenha garantido o contexto de sua obtenção? Estes dados podem representar um risco a privacidade dos indivíduos ou instituições neles referenciados de alguma forma? Como as partes de sua estrutura lógica serão interligadas e como serão mantidas as interligações com outros conjuntos de dados? Como garantir que os elementos que sustentam a sua qualidade sejam mantidos? Tem-se o direito de armazenar estes dados? Todos os aspectos que podem contribuir para sua encontrabilidade estão sendo armazenados? Todos os fatores para sua utilização ao longo do tempo estão sendo mantidos?

Portanto nesta fase percebe um conjunto de planejamentos e ações que requerem um conhecimento mais profundo da Ciência da Computação, mas

que ainda apresenta forte potencial de participação para a Ciência da Informação. Já o usuário fica um pouco mais distante, participando mais ativamente apenas da validação dos modelos de estruturas definidos para os dados.

Após esta fase pode-se chegar a um momento em que se decide que estes dados já não são necessários ou que não devem ser mantidos, o que leva ao seu descarte, mais uma vez referenciando-se a outra fase a ser discutida. No entanto, o mais comum será buscar alternativas que permitam o acesso e uso destes dados. Começa, assim, uma nova fase em que preocupações e esforços são voltados para estes dados possam ser encontrados, acessados e interpretados. Trata-se de uma fase em que o objetivo passa então a ser a viabilização da **recuperação** destes dados.

Dentre outras questões fundamentais desta fase pode-se destacar: quais dos dados armazenados serão disponibilizados? Existe algum público alvo específico? Já se tem uma necessidade que se pretende atender ou pelo menos um escopo do que se pretende disponibilizar? O acesso será feito diretamente a base em que se encontra armazenado ou será necessário retornar a fase de armazenamento para definição de novas estruturas de armazenamento específicas para recuperação? Com que frequência os dados serão atualizados para disponibilização? Quem poderá acessar estes dados? Durante o processo de recuperação quais são os riscos à privacidade dos indivíduos ou entidades referenciados pelos conteúdos recuperados? Como explicitar e operacionalizar a integração entre as diversas estruturas dos dados e destes com outros conjuntos de dados? Como explicitar e garantir os elementos que sustentam a qualidade dos dados que estão sendo disponibilizados? Têm-se o direito de disponibilizar estes dados? Como viabilizar que estes dados sejam encontrados, acessados e passíveis de interpretação (preferencialmente, e em muitos casos obrigatoriamente, por máquinas)? Os processos e procedimentos de recuperação estão estáveis o suficiente para que permaneçam polimorficamente utilizáveis ao longo do tempo?

Nesta fase fica explícita a necessidade de conhecimentos advindos da Ciência da Informação, instrumentalizados pela Ciência da Computação e

preferencialmente contando com aqueles que detenham conhecimento sobre o público alvo ou em potencial bem como das necessidades previstas, por mais amplas que sejam suas definições.

Em determinados pontos desta fase também se tem a possibilidade de se identificar que os dados já não são necessários ou que devem ser excluídos da base, o que leva a outra fase que responde pela limpeza ou simplesmente desativação de parte dos dados. Nesta fase, identificada como fase de **descarte** tem-se então a eliminação de parte dos dados que pode ocorrer em bloco, horizontalmente ou verticalmente. Em bloco seria a exclusão de subconjuntos inteiros de dados identificados como entidades (SANTOS; SANT'ANA, 2013). No caso de eliminações horizontais teríamos a eliminação de registros (elementos da estrutura entidade) por meio de filtros específicos ou de informações relacionadas às datas a que se relacionam. Verticalmente seria a eliminação de elementos estruturais das entidades que definem seus atributos, o que remete a definição de dado como sendo definido pela tríade <e,a,v> (SANTOS; SANT'ANA, 2013). Exemplificando, uma eliminação em bloco poderia ser a exclusão de uma entidade que contém os dados de produtos (inteiramente apagada). Já para uma eliminação horizontal relativa a esta mesma entidade podemos imaginar a eliminação de produtos que tenham sido cadastrados a mais de x anos, e para uma eliminação vertical seria a eliminação de um elemento estrutural desta entidade e que, portanto, identificaria um de seus atributos, como por exemplo, peso e, portanto, sua eliminação apagaria o atributo 'peso' de todos os itens cadastrados nesta entidade.

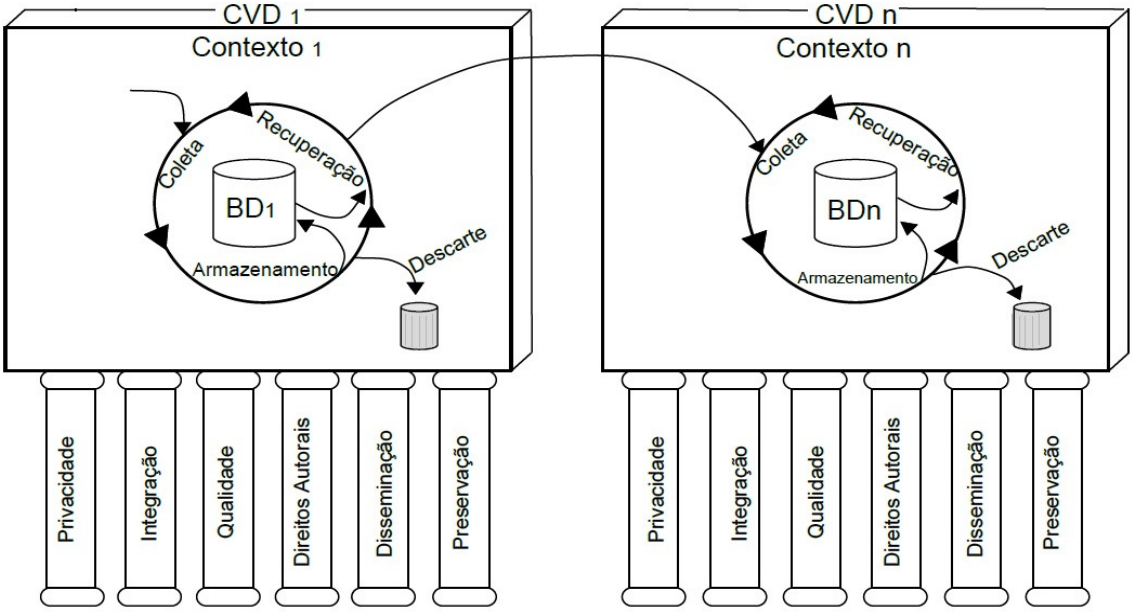
Para esta fase também são necessários conhecimentos específicos para os esforços de planejamento e execução, sendo que a Ciência da Informação pode desempenhar um importante papel, mais uma vez instrumentalizada pela Ciência da Computação e contando ainda com o aval e acompanhamento dos usuários envolvidos.

Dentre outras questões fundamentais desta fase pode-se destacar: Quais dados já não são mais necessários? Os dados a serem descartados foram persistidos? Em quais suportes? Estes dados estão replicados em outras bases? Como garantir e explicitar que estes dados foram realmente excluídos e

não simplesmente ocultos? A eliminação destes dados não prejudicará a integridade ou interligação de outros dados? O descarte destes dados não prejudicará a qualidade do conjunto de dados como um todo? Têm-se o direito de excluir este dado? Ao eliminar estes dados qual o impacto em sua encontrabilidade e acesso? Para o descarte foi considerada a necessidade de preservação em seus diversos aspectos?

Aponta-se, assim, para a existência de quatro fases e de fatores que permeiam (estão presentes em) todas elas, que são: privacidade, integração, qualidade, direitos autorais, disseminação e preservação, conforme descrito na figura 1.

Figura 1 - Ciclo de Vida dos Dados para Ciência da Informação – (CVD–CI)



Fonte: Adaptado de Sant'Ana (2013).

Desta forma podemos explicitar as fases de Coleta, Armazenamento, Recuperação e Descarte e os fatores que estão presentes em cada uma destas fases: Privacidade, Integração, Qualidade, Direitos Autorais, Disseminação e, Preservação (SANT'ANA, 2013). Para explicitar cada uma destas fases e fatores do CVD, passamos a analisar cada uma delas em relação ao contexto de um projeto de repositório de dados.

Coleta

Em um primeiro momento, depara-se com a fase de obtenção dos dados que pode ser identificada como aquela em que ocorrem: a definição das necessidades informacionais que irão nortear as escolhas e definições sobre quais dados são necessários; são estabelecidas estratégias sobre como localizar e avaliar estes dados; são escolhidos os mecanismos que serão utilizados para sua obtenção, e; são elaboradas as metodologias e ferramentas necessárias para consecução destes dados.

Nesta fase, denominada no Ciclo de Vida dos Dados como Fase de Coleta, a estruturação de um Repositório de Dados deve considerar a difícil tarefa de identificar fontes de dados que possam ser úteis para que os usuários possam ter atendidas suas necessidades.

A coleta pode ser caracterizada como um projeto ou como um processo. Existem casos em que a coleta ocorre a partir de fonte de dados que permite a aquisição constante dos mesmos, o que geralmente configura situação em que estes dados correspondem a informações que representam situações dinâmicas constituindo um processo, que pode ser contínuo, de fornecimento destes dados. Neste contexto, entre outros aspectos a serem considerados, emerge a questão da definição da cadência da coleta, por meio da identificação do tempo entre as tomadas das medidas ou obtenção dos valores. Um intervalo menor gerará uma maior precisão, gerando, no entanto, um volume maior de dados coletados. Já um intervalo maior gera volumes menores em detrimento a granularidade. Este tipo de coleta configura um processo que acaba por coexistir com as demais fases.

Já nos casos em que a coleta é pontual, cada procedimento de coleta pode ter suas configurações próprias e o estabelecimento dos metadados deve levar em conta as características de cada coleta, principalmente do registro do momento em que esta ocorreu, configurando as características de um projeto, com definição clara de início e de término. Situações como esta podem, ainda, indicar necessidade de análise de novas realizações de coleta, em um movimento cíclico de atualização dos dados coletados, o que o aproxima da coleta contínua, porém com características próprias, em que são definidos blocos, subconjuntos de dados, dentro do escopo dos dados coletados.

Neste contexto, pode-se a analisar os fatores envolvidos nesta fase, a começar pela **privacidade**, que nesta área ganha contornos de grande destaque. Se faz necessário identificar, nas fontes utilizadas, aspectos que possam configurar quebra de privacidade de pessoas ou instituições relacionadas aos dados que estão sendo coletados o que poderia resultar em um passivo futuro a partir da base de dados obtida, comprometendo as próximas fases do ciclo de vida.

Na coleta também é importante ter pronta a definição de requisitos sobre a base de dados que se pretende obter como um todo e os relacionamentos necessários para que esta base possa ser vinculada a outras bases, proporcionando um resultado que remete à questão do valor do todo que tende a ser maior que a soma das partes quando estas partes estão devidamente integradas. Assim, a **integração** deve ser preocupação da fase de coleta por meio da identificação e validação dos atributos que serão responsáveis pela identificação unívoca de cada registro (chave candidata ou primária) e seus correspondentes nas outras entidades (chaves estrangeiras) para que a integração possa ser garantida.

Uma característica essencial de repositórios de dados é a definição e garantia de elementos que permitam a percepção da qualidade dos dados coletados, portanto, elementos como a procedência, mecanismos de coleta e garantias de integridade física e lógica representam apenas alguns dos aspectos a serem considerados. A confiabilidade dos dados é condição *sine qua non* para que um dado seja útil.

No momento da coleta deve-se manter em foco a questão do responsável pela fonte de dados que se pretende utilizar para que não venha a ser desrespeitado o **direito autoral** vinculado aos dados que serão obtidos, considerando, sempre, que podem existir empresas e recursos envolvidos no desenvolvimento de soluções para o tema cuja coleta foi alvo, e que, independente de discussões éticas sobre o direito de acesso ou não a determinados recursos informacionais, deve-se, em última instância, respeitar o arcabouço jurídico que sustenta a legalidade deste acesso.

Deve-se consultar sempre as informações sobre direito de acesso aos dados desejados e suas nuances como a questão de resultados derivados ou

de vinculação financeira de uso futuro de resultados produzidos a partir deles e ainda a autorização de alteração e de obrigatoriedade de citação de fonte. O maior volume possível destas informações deve ser devidamente registrado no âmbito do próprio repositório, ampliando sua segurança jurídica e ainda de seus responsáveis.

O eventual acesso futuro, e, portanto, a **disseminação** destes dados, deve ser considerada já desde a fase de coleta de tal forma que a viabilidade de uma maior encontrabilidade e acesso seja possível, exigindo que informações (atributos) que mesmo que não estejam ligados diretamente a necessidade atual sejam incluídas no planejamento da estrutura de obtenção para que seja possível, por exemplo, identificar elementos contextuais dos dados que possam favorecer sua localização e interpretação na fase de recuperação. Isso pode gerar, inclusive, a ampliação do volume de dados obtidos e de aporte de informações sobre a própria coleta como parte das estruturas a serem preenchidas com os dados coletados.

A **preservação** dos dados coletados e que eventualmente venham a ser armazenados, também podem exigir que dados adicionais sejam incluídos nos pré-requisitos definidos para a coleta, proporcionando que estes dados possam ser identificados de forma mais ampla e incorporando, inclusive, informações sobre eventuais características de dispositivos que tenham sido utilizados como fonte dos dados, permitindo que estes dados sejam não só preservados mas também utilizados, mesmo após inevitáveis alterações em suas estruturas e constituições semânticas advindas de evolução nos dispositivos, tais como aumento de acurácia e ou precisão, com eventuais, e muito prováveis, melhorias nos níveis de granularidade de informações.

Armazenamento

Uma vez coletados os dados, surge o potencial uso posterior dos mesmos, ou seja, a possibilidade de que estes dados possam ser utilizados em novos processos de análise direta ou por meio de interação com outras bases de dados, o que leva a necessidade de estruturação de metodologias e ações relacionadas ao que na Ciência da Computação é definido como persistência

dos dados (RUMBAUGH *et al*., 1994, p. 429) e que no CVD é definido como fase de Armazenamento.

Nesta fase tem-se um enfoque mais tecnológico e se definem aspectos que garantem a reutilização destes dados, por meio de especificações físicas e lógicas sobre como os dados serão registrados em um suporte. Algumas das definições necessárias nesta fase são:

i. Qual o conjunto de variáveis que receberá os conteúdos (valores) obtidos na fase de coleta e para cada uma destas variáveis será necessário definir quais são suas especificações, tais como seu tipo (se será um conteúdo compostos por um valor numérico, por um conjunto de caracteres, por um valor lógico, por um conjunto de *bytes* que permita outros conteúdos tais como som, imagens, vídeos), tamanho, formato e ainda especificações semânticas tais como: qual sua unidade de medida, grau de precisão e tudo mais que puder facilitar a interpretação futura deste dado;

ii. Com que estrutura, ou seja, o conjunto de variáveis definida no item i precisa ainda ser organizado em subconjuntos definidos de acordo com a semântica que os vincula a um elemento ou conceito do mundo real e que os define, assim, tem-se como exemplo um conjunto de variáveis que podem ser atribuídas diretamente a um produto e que, portanto, comporiam uma entidade chamada Produto com variáveis como: descrição, data de cadastro, peso, número de identificação que possa identificá-la e assim por diante. Considerando as definições i e ii, tem-se então a estrutura semântica mínima para interpretação e uso de um dado: <e,a,v>, ou seja, a qual entidade este atributo (variável) pertence e qual o seu valor (SANTOS; SANT'ANA, 2013). Portanto, na fase de planejamento de um repositório de dados será necessário identificar para cada um de seus conjuntos de dados quais e como estão estruturados em termos de entidades e atributos, e, também, como seus valores estão registrados permitindo que para cada dado armazenado seja possível interpretar sua estrutura básica <e,a,v>;

iii. Quem poderá acessar os dados armazenados é outra questão fundamental, principalmente quando se trata de conteúdos com grande possibilidade de serem identificados como dados sensíveis, e que, portanto, suscitam questões sérias sobre a privacidade no acesso a estes dados. Nesta

questão, a estrutura básica de um dado permite perceber parte da complexidade da definição de regras de acesso a estes dados, já que estas definições devem levar em conta que não se trata de um simples item de informação, mas que deverá ser pensada a estrutura como um todo, ou seja, deverá ser pensada em termos de entidade, atributo e valor. Autorizações de acesso deverão levar em conta não somente quais entidades poderão ser acessadas mas, também, quais atributos e ainda como os valores serão tratados para o acesso. Com relação aos valores, pode-se usá-los para se restringir o acesso de um registro em função do valor de um determinado atributo, por exemplo, restringindo o acesso a dados sobre produtos somente aos representantes que estão envolvidos com a sua comercialização ou acompanhamento. Pode-se, ainda, estabelecer níveis de anonimização para os dados por meio de tratamento de determinados valores de atributos considerados como identificadores ou semi-identificadores por meio de generalização ou mesmo supressão (SAMARATI; SWEENEY, 1998);

iv. Como serão acessados, seja de forma direta ou por meio de um Sistema Gerenciador de Banco de Dadas (SGBD), que interfere inclusive no formato físico que será adotado para registro destes dados. No caso de adoção de um SGBD o mais indicado é que seja mantido o próprio formato utilizado pelo sistema, o que faz com que os dados fiquem sob a gestão destes e fora do acesso direto aos conteúdos originais, o que, se por um lado reduz a interação mais direta, amplia a segurança tanto física quanto lógica. Já para o caso de se optar por um padrão que permita o acesso direto, poderá ser adotado desde o padrão de formatação baseada em semântica posicional até formatos mais elaborados e que incluem a semântica que define as definições das entidades e dos atributos por meio de metadados incorporados aos conteúdos o que pode propiciar a interpretação e uso destes dados, inclusive, de forma automatizada;

v. em que formato ou padrão, definição derivada da definição iv tem como principal objetivo permitir que a semântica das entidades e dos atributos possam ser disponibilizadas em conjunto aos conteúdos. Como exemplos desta definição pode-se citar escolhas como a do formato *Comma Separated Values* ou CSV que é muito simples e permite acesso facilitado por meio de

uma simples planilha. Seu formato é baseado em uma planilha em que cada linha do arquivo, delimitada por uma quebra de linha (*carriege return e line feed* - CRLF), representa uma linha da planilha e o conteúdo de cada coluna é separado por um caractere escolhido, na maioria das vezes opta-se pelo uso da vírgula como o próprio nome indica. Opcionalmente, pode-se adotar a primeira linha como cabeçalho, o que permite que se tenha, pelo menos, o rótulo de cada uma das colunas (IETF, 2005);

vi. Onde estarão armazenados é outra definição que vem ganhando importância crescente já que em um movimento cíclico percebe-se um retorno ao modelo de armazenamento utilizado nos primórdios da computação digital em que a dificuldade de armazenamento levava a uma centralização do armazenamento em dispositivos de grande porte. Com a disseminação dos dispositivos de baixo custo e com capacidades cada vez maiores de armazenamento, os conteúdos passaram a se fragmentar em cada um dos equipamentos que deles necessitavam, trazendo grandes vantagens de desempenho, mas agregando dificuldade de interoperabilidade. Hoje, com o advento da interconectividade massiva, observa-se uma crescente tendência à utilização de dispositivos de acesso cada vez mais focados na tarefa de interface e abdicando da responsabilidade pelo armazenamento. Os dados tendem, então, a serem armazenados em algum lugar da rede, o que vem gerando um processo de centralização destes dados nos próprios provedores de serviços de informação.

Com base nos objetivos e definições descritos, passa-se a analisar os fatores envolvidos nesta fase e a **privacidade** esta fortemente relacionada ao item iii já que por meio desta definição serão identificados aqueles que poderão ter acesso a estes dados, não só para consulta, mas, também, para inclusão, alteração e até mesmo exclusão de informações.

Com relação ao item iv uma das grandes vantagens da adoção de SGBDs está justamente na possibilidade de se definir quem poderá ter acesso aos dados, sendo que isso ainda pode ser feito por meio de definições de papéis de usuário, onde se define não somente a identificação de usuários mas grupos destes usuários que podem estar relacionados as atividades que este grupo desenvolve, o que permite que sejam definidas atribuições de acesso

para um determinado papel e posteriormente cada usuário pode então ser identificado a um ou mais papéis o que define seu escopo de acesso aos conteúdos de uma base de dados.

No item vi que define "onde" os dados serão armazenados também pode gerar uma série de questões relacionadas à privacidade, sendo que uma base armazenada localmente e desconectada da rede pode estar muito mais segura com relação a acessos ou usos indevidos do que uma base de dados que esteja armazenada em um servidor de dados conectado a internet, muitas das vezes, sob a responsabilidade de terceiros. Nos casos em que se têm muitos dados sensíveis esta questão é primordial, levando, geralmente, a uma grande fragmentação do armazenamento destes dados.

Uma preocupação que deriva destes fatores é: como garantir a privacidade no acesso a dados que estejam armazenados em dispositivos intramuros de uma empresa?

A possibilidade de **integração** dos dados armazenados vai depender em grande parte das definições descritas nos itens iv e v. Na definição de como serão acessados, a escolha pela adoção de um SGBD fará com que seja criada uma camada de proteção e de interação que necessariamente deverá ser articulada para que possa ocorrer integração. Já no caso de adoção de um formato aberto que permita o acesso direto pode facilitar o acesso, mas traz o inconveniente de necessitar de uma forte carga semântica nos dados que permita a identificação e interpretação das entidades e atributos daquela base de dados.

As definições dos itens i e ii também interferem nesta questão, já que uma vez acessados estes dados deverão estar articulados entre si por meio de relacionamentos possíveis de serem identificados pelos próprios algoritmos o que leva a necessidade de conjuntos bem definidos de identificadores, não só do próprio registro, mas também viabilizando sua relação com outras entidades, tudo isso obtido por meio de definições bem feitas de chaves primárias (identificação do próprio registro) e de chaves estrangeiras (identificação de chaves primárias de outras entidades ou da mesma entidade em relacionamentos dentro da própria entidade). Quando se trata de dados sensíveis a questão se agrava já que a definição bem feita de identificadores e

até semi-identificadores pode ampliar, e muito, as possibilidades de uma ação e ataque, que, por meio da integração de bases que não ferem diretamente a privacidade formando novas bases de dados com grande potencial de quebra da privacidade.

Quanto à **qualidade** dos dados, as definições sobre seu armazenamento são fundamentais para garantir que estes dados mantenham sua integridade física e lógica. Um dado armazenado de tal forma que não dependa de um SGBD para o acesso terá muito mais possibilidade de vir a ter um problema com relação a sua integridade, seja por problemas físicos ou por acessos indevidos que podem ocorrer por má fé ou mesmo por erros operacionais. Para os casos de dados sensíveis a qualidade dos dados ganha dimensão extra já que decisões tomadas a partir destes podem gerar consequências irreparáveis.

Ao se armazenar dados, uma das preocupações deve ser com relação aos **direitos autorais** vinculados a fonte da qual os dados foram obtidos, buscando-se registrar, também, estas informações para que se mantenha a segurança institucional daqueles que respondem pelos dispositivos de armazenamento. Existem áreas que apresentam forte presença do estado em sua gestão o que pode favorecer o acesso e futuro armazenamento destes dados, mas também deve ser considerada a massiva presença de empresas que realizam investimentos em pesquisa e desenvolvimento e que, portanto, esperam por resultados financeiros de seus investimentos e, assim, buscam proteger seus ativos informacionais. Independente de discussões éticas a este respeito deve-se ter em mente que um dado armazenado deve conter em seus metadados informações sobre sua origem e sobre os direitos relacionados a ela.

Cabe destacar, ainda, que mesmo quando se trata de armazenamento de tratamento de outras bases de dados deve-se considerar a questão de trabalho derivado que mesmo não sendo uma cópia dos dados originais o conteúdo armazenado pode ser resultado de acesso a determinadas bases e que só seriam possíveis por meio deste acesso, deverão conter as informações sobre os dados que os originaram.

Espera-se que os dados armazenados proporcionem acesso futuro, assim, a **disseminação** é uma das preocupações que também se apresentam na fase de armazenamento. É necessário prever meios que permitam que estes dados estejam acessíveis e ainda incorporar semântica para que sejam passíveis de interpretação, preferencialmente automatizada. É necessário possibilitar que a base de dados contenha elementos que permitam e facilitem, também, a sua localização. Quando se tratar de dados sensíveis, uma questão que vem à tona quando se trata da disseminação e mais uma vez a privacidade, e no momento de criar meios para que estes dados sejam encontráveis devem ser planejadas alternativas de proteção previstas no armazenamento.

Ainda com o enfoque de uso futuro, deve-se, no armazenamento, prever elementos que propiciem a **preservação** destes dados, de tal forma que se possa interpretá-los no futuro, independente do acesso aos seus responsáveis ou aos dispositivos originais que o armazenaram o que leva a elaboração de uma estratégia de execução de processos de atualização tecnológica e de verificação de integridade física e lógica.

Quando se trata de preservar dados no contexto do *Big Data*, deve-se levar em conta não somente os aspectos comuns ao processo de preservação, mas, também, fatores, como por exemplo, a vasta gama de formatos e de variedades de fontes de dados, bem como a diversidade de dispositivos de coleta que, ainda, apresentam a constante evolução como agravante na questão da manutenção das informações sobre sua obtenção.

No que diz respeito à preservação de aspectos semânticos a questão não é menos complexa, com a necessidade de registro não só dos conteúdos, mas, também, de elementos que permitam a interpretação de vocabulários específicos, regionais e de outras épocas. Tudo isso deve estar de alguma forma armazenado e relacionado ao conteúdo que se pretende preservar.

Recuperação

Uma vez que os dados tenham sido coletados e estejam armazenados pode-se proporcionar uma nova fase que seria aquela em que, tomando como foco os dados, passa-se a tornar estes dados disponíveis para acesso e uso.

Assim, têm-se a fase de recuperação dos dados coletados e armazenados e as estratégias e ações passam a ser avaliadas a partir do ponto de vista do responsável por sua manutenção e não daqueles que acessarão estes dados (para estes a fase é a de coleta). Isto não implica que características dos que recuperarão os dados não são consideradas, muito pelo contrário, este é o objetivo, mas o foco, agora, é de quem está propiciando esta recuperação já que está se levando em consideração como foco, a base de dados.

Nesta fase preocupa-se com meios que ampliem os níveis de utilização destes dados, seja por ampliação das possibilidades de acesso via cópia ou obtenção de conjuntos para análise seja por meio da disponibilização de recursos de visualização destes dados. Este é um tema bastante amplo e fugiria ao escopo deste texto esgotá-lo, mas podemos abrir algumas reflexões sobre os fatores envolvidos nesta fase.

No que diz respeito à **privacidade**, quando se pensa em meios de recuperação para os dados, vale lembrar que devem ser considerados os envolvidos com os conteúdos a serem disponibilizados, identificando estruturas e possíveis usuários, lembrando-se sempre que, mesmo em casos em que os dados não se mostrem sensíveis, deve-se prever a possibilidade de vinculação destes dados com outros dados, o que pode propiciar um ataque. Quando se tratar de dados sensíveis, a privacidade deve ser tema central das preocupações na disponibilização de dados para recuperação. O nível de anonimização nestes casos deve ser o mais elevado possível, mesmo tendo-se em mente a deterioração do nível de utilidade de uma base de dados diretamente relacionado ao grau de anonimização imposto a ela.

Para que se possa obter um bom nível de uso, a partir de dados armazenados, estes precisam apresentar um grau de **integração** que propicie análises de entidades distintas, mas integradas, de forma a comporem um todo que pode representar um valor de uso maior que a soma dos valores de uso das entidades individualmente.

Viabilizadas por um bom planejamento e cuidadosa execução nas fases de coleta e armazenamento, na fase de recuperação estes conjuntos de dados formados, na maioria das vezes, por muitas entidades, terão seu acesso unificado e descrito como uma única entidade, o que facilita ao usuário obter os

resultados esperados. Mesmo considerando a necessidade da manutenção da privacidade, um dos grandes problemas é que o uso destes dados requer uma forte integração dos mesmos, já que o histórico de uma entidade pode ser de fundamental importância para a realização de um diagnóstico ou para a definição de análises definidas como horizontais destes dados.

Assim como na coleta e no armazenamento, também na recuperação a **qualidade** é um fator chave no ciclo de vida dos dados. Os recursos disponibilizados, seja de acesso para *download* seja para visualização de dados, devem refletir os mesmos aspectos, no entanto, com especial atenção à interação do usuário. Devem ser considerados aspectos como a arquitetura da informação na elaboração dos recursos de recuperação e, ainda, elementos que ampliem a usabilidade e a acessibilidade, evitando possíveis erros derivados da própria interface.

Com relação aos **direitos autorais** na fase de recuperação, deve-se deixar explícito quem tem permissão de usar e como os dados podem ser utilizados a partir desta recuperação, tornando mais fácil e segura sua utilização e replicação.

Para a **disseminação** na fase de recuperação é necessário dotar os dados coletados e armazenados com elementos que permitam que os mesmos sejam encontrados por aqueles que irão utilizá-los, não bastando a simples possibilidade de acesso. São necessárias estratégias que permitam sua localização, não somente para acesso pelos próprios recursos de visualização de seus detentores, mas, também, por mecanismos automáticos que possam não só encontrá-los como ainda acessá-los em processos de coleta.

Quando se trata de dados sensíveis, os usuários tendem a ter identificação forte e com direitos restritos de acesso, mas, mesmo assim, estes usuários precisam receber a informação de que aquele dado está disponível. Devem estar disponíveis, também, todas as informações sobre como usá-los, os aspectos semânticos envolvidos e, ainda, limitações de acesso, de tal forma que tudo isso possa ser identificado ainda no momento da localização, para facilitar a decisão sobre seu uso ou não.

A **preservação** na fase de recuperação está diretamente ligada a questões de sua interpretação, principalmente quando relacionada ao fator

tempo, ou seja, uma interpretação realizada em um determinado momento deve ter a possibilidade de ser a mesma realizada em outro momento, desde que mantidos os critérios e objetivos originais, assim, tanto os recursos de preparação e seleção para *download* como os recursos de visualização devem manter um rígido controle sobre seus algoritmos e mecanismos de interação para não gerar resultados distintos a partir de uma base ao longo do tempo.

Esta não é uma tarefa fácil, principalmente quando se considera a constante pressão por atualização e modificação que estes recursos sofrem durante sua existência. Quando se trata de dados sensíveis esta questão se torna bastante preocupante já que uma pesquisa realizada em momentos distintos podem gerar diferentes resultados levando a conclusões sobre causas ou consequências de um determinado objeto de estudo o que pode levar a condução de decisões equivocadas.

Descarte

Uma vez concluídas as reflexões sobre as fases de coleta, armazenamento e recuperação, poderia se supor que o ciclo de vida dos dados esta completo, principalmente em um momento em que o limite para o volume de dados parece cada vez mais alto, mas não é o que ocorre. Vivencia-se um momento em que definições como a de *Big Data* identificam um cenário em que o volume de dados que a disposição é cada vez maior, ultrapassando a capacidade de interpretação e mesmo de armazenamento eficiente. Cabe refletir sobre uma fase em que ocorre o descarte de dados que não são mais necessários ou que estejam acima da capacidade de tratá-los com eficiência para o sistema como um todo.

Quando se trata de questões relacionadas à fase de descarte, não é trivial a vinculação desta com a questão da **privacidade**, mas ela é bastante premente nesta fase e merece atenção. Um indivíduo deve ter o direito ou pode vir a ter a necessidade de ter seus dados retirados de uma determinada base e garantir o que poderíamos identificar com o conceito do direito ao esquecimento.

Mas esta não é uma tarefa fácil já que o acesso aos dados será sempre mediado de alguma forma pelos seus detentores e limitações ao acesso direto

podem apresentar cenários incompletos sobre os dados que permanecem armazenados, podendo gerar a percepção de que um determinado dado foi excluído quando na verdade ele foi apenas identificado como não acessível para visualização e, portanto, permanecendo registrado para análises intramuros e fora do alcance prático de instâncias de acompanhamento e controle.

Outra questão relacionada ao direito ao esquecimento em uma base de dados pode estar relacionada à existência de cópias destes dados que podem estar armazenadas em locais distintos, longe da possibilidade de controle ou acompanhamento por parte daqueles que ali tem seus dados registrados.

Já com relação à **integração** dos dados no momento do descarte, a relação é mais explícita e suscita uma série de preocupações diretas, pois, um determinado registro excluído de uma base pode causar a degeneração de relacionamentos entre bases distintas levando a uma degradação do valor de uso da base como um todo.

Vale lembrar ainda que os dados que estão sendo descartados podem já ter sido utilizados por terceiros ou proporcionado conteúdos derivados e que por sua exclusão seriam todos afetados.

Com relação a questão da **qualidade**, que nesta fase se relaciona de forma muito direta com o fator integração, tem, ainda, a fase de descarte, o agravante da destruição de conceito de conjunto retirando, eventualmente, a possibilidade de análises comparativas ao longo do tempo ou a partir de contextos diferentes, e pior, podendo levar a conclusões equivocadas a partir de suas análises.

Por isso, processos de descarte devem manter informações registradas com o maior detalhe possível sobre os processos de eliminação, e de tal forma, que em consultas futuras que possam ter, de alguma forma, relação com os contextos afetados, estas informações sejam apresentadas para que as eliminações sejam consideradas em novas análises.

Questões de **direito autoral** na fase de descarte podem ser relacionadas a principalmente a manutenção de informações sobre autoria mesmo depois de descartadas já que estes dados podem ter sido utilizados por terceiros que em determinado momento precisariam justificar ou identificar a

origem de seus dados e que após o descarte perderiam também as informações sobre autoria e seus registros, assim, deve-se manter de alguma forma, informações sobre dados que já foram disponibilizados e que de alguma forma podem ter sido utilizados por terceiros para não gerar insegurança legal em possíveis trabalhos derivados ou mesmo de utilização direta e ou apenas referenciada.

Sobre a **disseminação** e sua relação com a fase de descarte, vale ressaltar o efeito *pipeline* principalmente em mecanismos de busca que permaneceriam com informações sobre o conteúdo de um determinado conjunto de dados quando na verdade estes dados já não estão disponíveis. Outra preocupação sobre este fator na fase de descarte refere-se ao risco de se perder elementos que, mesmo não estando relacionados ao foco do conjunto de dados em questão, podem representar elementos chave para encontrabilidade do conjunto.

A **preservação** apresenta correlação direta com a fase de descarte, relacionando-se inclusive com os outros fatores. A preservação deve ser buscada mesmo quando o dado não parece ser mais útil, já que sempre podem surgir novas necessidades, não previstas, que venham a requerer os dados eliminados.

Em função de custos de armazenamento cada vez menores, têm-se a possibilidade de se manter cópias de dados que, por motivo de eficiência do sistema, tenham de ser eliminados. Vale lembrar que o armazenamento de uma cópia de dados excluídos, muitas das vezes em formatos e estruturas diferentes das originais, configura o início de um novo ciclo de vida dos dados, já que se tratará de uma nova base de dados com suas características e objetivos específicos.

3 CONCLUSÕES

A temática de acesso a dados é muito ampla e não seria viável propor seu esgotamento em um texto, mas buscou-se, aqui, apresentar reflexões que pudessem abarcar todo o contexto com o qual estão envolvidos o acesso, o uso e a manutenção de dados.

O ciclo de vida dos dados descrito neste artigo não tem como objetivo ser um fim e sim uma forma de proporcionar uma estrutura que suporte os esforços, estudos e ações realizadas para obtenção, manutenção e uso de dados, tornando possível aproximar elementos semelhantes e distribuir teorias e metodologias em função de seu escopo seja por fase ou por fatores identificados no modelo de ciclo de vida dos dados.

Destaca-se o papel da Ciência da Informação (CI) no estudo e na proposta de soluções para todo o processo de acesso a dados, oferecendo um arcabouço teórico que possa contribuir na construção deste novo cenário, em conjunto com outras áreas como a Ciência da Computação e a Matemática. Destaca-se que os estudos e pesquisas em andamento na CI apresentam maior aderência nas fases de coleta e recuperação, sendo que as fases de armazenamento e descarte, por sua natureza mais técnica dependem mais fortemente da Ciência da Computação, o que não diminui a importância de contribuições da CI para estas fases. A CI pode, portanto, complementar seu papel na proposta de novos caminhos para que os usuários, de forma mais democrática e aberta possam acessar e utilizar dados, reduzindo a assimetria informacional que pode surgir entre os que os detêm e os que dele precisam.

Vale lembrar, ainda, a importância da temática dos metadados na análise e implementação de um ciclo de vida dos dados, já que todas estas definições deverão estar registradas, de alguma forma, em outro conjunto de dados (mesmo que faça parte da mesma base) e que, portanto, será responsável pelo registro dos dados sobre os dados, representando, assim, a estrutura dos dados coletados, armazenados e passíveis de recuperação e

REFERÊNCIAS

AKERLOF, George, A. The market for "lemons": quality uncertainly and the market mechanism. **The Quarterly Journal of Economics**, Cambridge, v. 84, n. 3, p. 488-500, Aug. 1970. Disponível em: <http://socsci2.ucsd.edu/~aronatas/project/academic/Akerlof%20on%20Lemons.pdf>. Acesso em: 10 fev. 2012.

CHEN, Ya-Ning, CHEN, Shu-Jiun, LIN, Simn C. A metadata lifecycle model for digital libraries: methodology and application for an evidence-based approach to library research. In: WORLD LIBRARY AND INFORMATION CONGRESS, 69., IFLA GENERAL CONFERENCE AND COUNCIL, 2003, Berlin. **Anais…** Berlin, 2003. Disponível em: <http://archive.ifla.org/IV/ifla69/papers/141e-Chen_Chen_Lin.pdf>. Acesso em: 10 maio 2013.

COAD, Peter; YOURDON, Edward. **Object oriented analysis**. New Jersey: Prentice-Hall, 1990.

DATA DOCUMENTATION INITIATIVE – DDI. **Structural reform group**: DDI Version 3.0 conceptual model. DDI Alliance. 2004. Disponível em: <http://libraries.mit.edu/guides/subjects/datamanagement/cycle.html>. Acesso em: 2 dez. 2012.

DATA OBSERVATION NETWORK FOR EARTH – DataONE. **Best practices**. Disponível em: <http://www.dataone.org/best-practices>. Acesso em: 10 maio 2013.

DAVENPORT, Thomas H. **Big data at work**: dispelling the myths, uncovering the opportunities. Harvard: Harvard Business School Publishing, 2014.

DEMARCO, Tom. **Structured analysis and system specification**. New Jersey: Prentice-Hall, 1979.

DIGITAL CURATION CENTER – DCC. **Curation lifecycle model**. Disponível em: <http://www.dcc.ac.uk/resources/curation-lifecycle-model>. Acesso em: 12 jan. 2013.

FERDERER, David A. **A data management life-cycle**. USGS Fact Sheet: 163-00. 2001. Disponível em: <http://www.usgs.gov/>. Acesso em: 10 maio 2013.

HUMPREY, Charles. **e-Science and the life cycle of research**. 2006. Disponível em: <http://www.usit.uio.no/om/organisasjon/uav/itf/saker/forskningsdata/bakgrunn/life-cycle.pdf>. Acesso em: 15 maio 2013.

INTERAGENCY WORKING GROUP ON DIGITAL DATA - IWGDD. **Harnessing the power of digital data for science and society**. 2009. Disponível em: <http://www.nitrd.gov/About/Harnessing_Power_ Web.pdf>. Acesso em: 10 maio 2013.

INTERNET ENGINEERING TASK FORCE - IETF. **Common Format and MIME Type for Comma-Separated Values (CSV) Files**. Request for comments RFC 4180. 2005. Disponível em: <https://tools.ietf.org/html/rfc4180>. Acessado em: 10 maio 2013.

MATERIAL DATA MANAGEMENT CONSORTIUM - MDMC. **The materials data lifecycle**. Disponível em: <http://www.mdmc.net/pages/lifecycle.htm>. Acesso em: 25 jan. 2013.

MCMENAMIN, Stephen M.; PALMER, John F. **Essencial system analysis**. Michigan: Yourdon Press, 1984.

PENNOCK, Maureen. **Digital curation: a life-cycle approach to managing and preserving usable digital information**. 2007. Disponível em: <http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf>. Acesso em: 25 jan. 2013.

RUMBAUGH, James et al. **Modelagem e projetos baseados em objetos**. Rio de Janeiro: Campus, 1994.

SAMARATI, Pierangela; SWEENEY, Latanya. Protecting privacy when disclosing information: kanonymity and its enforcement through generalization and suppression. 1998. Disponível em: <https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf>. Acesso em: 20 jan. 2015.

SANT'ANA, Ricardo Cesar Gonçalves. Ciclo de vida dos dados e o papel da ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais...** Florianópolis, 2013. Disponível em: <http://enancib.sites.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319>. Acesso em: 5 maio 2014.

SANTOS, Plácida L. V. Amorim da Costa; SANT'ANA, Ricardo César Gonçalves. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, Brasília, v. 42, p. 199-209, 2013.

STEVENS, Wayne P.; MYERS, Glenford James; CONSTANTINE, Larry L. Strutured design. **IBM System Journal**, New York, v. 13, n. 2, p. 115-139, 1974.

UNIVERSITY COLLEGE LONDON - UCL. **MRC Centre of Epidemiology for Child Health**. 2012. Disponível em: <http://www.ucl.ac.uk/ich/research-ich/mrccech/data>. Acesso em: 15 mar. 2013.

VANDERBILT UNIVERSITY MEDICAL CENTER – VUMC. **Strategic plan for VUMC informatics**. 2005. Disponível em: <https://ncs.mc.vanderbilt.edu/Data/NonSecure/IC_Strategic_Plan_9-12-05.pdf>. Acesso em: 13 jan. 2013.
YOURDON, Edward. **Modern structured analysis**. New Jersey: Prentice Hall, 1989.

YOURDON, Edward; CONSTANTINE, Larry L. **Structured design**. New Jersey: Prentice Hall, 1978.

Title
Data life cycle: a perspective from the Information Science

Abstract
Introduction: Access and use of data as a key factor has been extended to several areas of knowledge of today's society. It's necessary to develop a new perspective that presents phases and factors involved in these processes, providing an initial analysis structure, allowing the efforts, skills and actions organization related to the data life cycle.
Purpose: This article is a proposal for a new look at the data life cycle, that assumes, as a central element, the data itself, supporting itself on the concepts and contributions that Information Science can provide, without giving up the reflections on the role of other key areas such as Computer Science.
Methodology: The methodological procedures consisted of bibliographic research and content analysis to describe the phases and factors related to the Data Life Cycle, developing reflections and considerations from context already consolidated in the development of systems that can corroborate the idea of centrality of data.
Results: The results describe the phases of: collect, storage, recovery and discard, permeated by transverse factors: privacy, integration, quality, copyright, dissemination and preservation, composing a Data Life Cycle.
Conclusions: The current context of the availability of large volumes of data, with great variety and at speeds that provide access in real time, setting the so-called Big Data that requires new concerns about access and use processes of data. The Information Science may offer a new approach, now centered in the data, and contribute to the optimization of Data Life Cycle as a whole, extending bridges between users and the data they need.

Keywords: Data Life Cycle. Information Science. Big Data.

Titulo
Ciclo de vida del datos: una perspectiva desde la Ciencia de la Información

Resumen
Introducción: El acceso y utilización de datos como un factor clave ha sido extendido a varias áreas de conocimiento de la sociedad de hoy. Es necesario desarrollar una nueva perspectiva que presenta fases y factores que intervienen en estos procesos, proporcionando una estructura de análisis inicial, permitiendo la organización de esfuerzos, habilidades y acciones relacionadas con el ciclo de vida de los datos.
Propósito: Este artículo es una propuesta para una nueva mirada sobre el ciclo de vida de los datos, que asume, como elemento central, los datos en sí, apoyándose en los conceptos y las contribuciones que la Ciencia de la Información puede proporcionar, sin renunciar a las reflexiones sobre el papel de otra área clave como la Informática.
Metodología: Los procedimientos metodológicos consistieron en la investigación bibliográfica y análisis de contenido para describir las fases y factores relacionados con el ciclo de vida de los datos, haciendo reflexiones y consideraciones desde los contextos ya consolidados en el proceso de desarrollo de sistemas que pueden corroborar con la idea de la centralidad de los datos.
Resultados: Los resultados describen las fases de: recoger, almacenamiento, recuperación y descarte, permeada por factores transversales: privacidad, integración, calidad, derechos de autor, difusión y preservación, que componen un ciclo de vida de los datos.
Conclusiones: El contexto actual de la disponibilidad de grandes volúmenes de datos, con gran variedad y velocidade que proporcionan acceso en tiempo real, el

establecimiento de la llamada Big Data, requiere nuevas preocupaciones acerca de los procesos de acceso y uso de los datos. La Ciencia de la Información puede ofrecer un nuevo enfoque, ahora centrada en los datos, y contribuir con la optimización del Ciclo de Vida de los Datos en su conjunto, haciendo puentes entre los usuarios y los datos que se necesitan.

Palabras clave: Ciclo de Vida de los Datos. Ciencia de la Información. Big Data.

Enviado em: 17.07.2016.
Aceito em: 20.11.2016.