Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação

Plácida L. V. Amorim da Costa Santos
Livre-docência pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) - Marília, SP - Brasil.
Doutora em Linguística pela Universidade de São Paulo (USP) - São Paulo, SP - Brasil.
Professora da Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) - Marília, SP - Brasil.
http://lattes.cnpq.br/7408791408049766
E-mail: placida@marilia.unesp.br

Ricardo César Gonçalves Sant'Ana
Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) - Marilia, SP - Brasil.
Professor da Universidade Estadual Paulista Júlio de Mesquita Filho - Tupã, SP - Brasil.
http://lattes.cnpq.br/1022660730972320
E-mail: ricardosantana@marilia.unesp.br

Recebido em: 15/08/2014. Aprovado em: 23/1/2015. Publicado em: 07/08/2015.

Resumo
O desenvolvimento acelerado de recursos tecnológicos e sua utilização nos processos de acesso a dados, de uso da informação e de geração de conhecimento solicitam da Ciência da Informação (CI) uma revisão e ampliação de seu quadro referencial sobre as possibilidades interpretativas e de análise dos conceitos sobre dado e granularidade. O conceito de dado precisa ser redimensionado, entendido e percebido como elemento básico nos fluxos informacionais, especialmente em um momento em que tanto se discute e se legisla sobre o seu acesso na administração de conteúdos, no favorecimento de sua visibilidade e na sua utilização e reutilização. O objetivo é iniciar a reflexão e o debate sobre os conceitos de dado, conjunto de dados e granularidade no domínio da CI. Os conceitos de dado e de granularidade são apresentados sob o enfoque da Informação e Tecnologia, no interior da CI, com reflexão sobre dados estruturados e não estruturados, apontando a relevância dos metadados na complementação da estrutura semântica mínima de um determinado dado e na análise de sua granularidade.

Palavras-chave: Dado. Granularidade. Informação e tecnologia.

Data granularity under the perspective of information and technology: an interpretation by Information Science

Abstract
The rapid development of technological resources and their use in data access, information use and knowledge generation processes demand from Information Science (IS) a revision and broadening of its referential framework on the possibilities of data and granularity concept interpretation and analysis. The concept of data needs to be scaled, understood and perceived as a basic information flow element, especially at a time when much discussion and legislation is done on content management data access, in favor of its visibility and its use and reuse. The goal is to initiate reflection and debate on the concept of data, data sets and granularity in the field of Information Science. The concepts of data and granularity are presented under the scope of Information and Technology, within Information Science, with a reflection on structured and unstructured data, indicating the relevance of metadata to complement the minimal semantic structure of a given data and in the analysis of its granularity.

Keywords: *Data. Granularity. Information and technology.*

Granularidad de datos bajo la perspectiva de información y tecnología: una interpretación de la Ciencia de la Información

Resumen

El rápido desarrollo de los recursos tecnológicos y su uso en el proceso de acceso a datos, el uso de la información y la generación de conocimiento solicitan de la Ciencia de la Información (CI) una revisión y ampliación de su marco referencial sobre las posibilidades de interpretación y análisis de los conceptos de datos y granularidad. El concepto de datos necesita ser ampliado, comprendido y percibido como un elemento básico en los flujos de información, sobre todo en un momento en que mucho se discute y legisla sobre el acceso a datos en la gestión de contenidos, a favor de su visibilidad y su utilización y reutilización. El objetivo es iniciar la reflexión y el debate sobre el concepto de datos, conjunto de datos y granularidad en el dominio de la CI. Los conceptos de datos y granularidad se presentan con un enfoque de Información y Tecnología, dentro de la CI, con una reflexión sobre datos estructurados y no estructurados, indicando la importancia de los metadatos para complementar la estructura semántica mínima de un determinado dato y la análisis de su granularidad.

Palabras clave: *Datos. Granularidad. Información y tecnología.*

INTRODUÇÃO

As indagações que levaram à elaboração deste artigo partiram da hipótese de que, em tempos de constante mutação, proporcionada pelo uso e o desenvolvimento acelerado de recursos tecnológicos empregados nos processos de acesso a dados, de uso da informação e de geração de conhecimento, a ciência da informação (CI) precisaria rever e/ou ampliar seu quadro referencial sobre as possibilidades interpretativas e de análise sobre os conceitos de dado e de granularidade.

As características de multi e interdisciplinaridade presentes na CI suscitam reflexões nas três vertentes preconizadas por Morin (1998) sobre ordem, separabilidade e lógica: 1) a discussão sem divisão, 2) a imprevisibilidade e 3) a oposição da racionalização fechada à racionalização aberta, "pensar a complexidade é respeitar a tessitura comum, o complexo que ela forma para além de suas partes" (MORIN, p. 15, 1998).

A natureza interdisciplinar da CI foi analisada por Saracevic (1995) ao relacioná-la com quatro áreas: ciência da computação, biblioteconomia, ciência cognitiva e comunicação, afirmando-a como uma ciência de resolução de problemas. Wersig (1993), em contraponto, a designa como interdisciplinar e ressalta seu aspecto social, sem desconsiderar o problema enfrentado por essa interdisciplinaridade, que revela a área como campo de estudo que "tem sido objeto de muitas disciplinas fragmentadas e, portanto, temos que lidar com todos esses itens fragmentados de uma natureza empírica ou teórica" (WERSIG, 1993, p. 235). Fato que conduz a uma necessidade fundamental de formular uma visão integradora do campo teórico e de aplicação a partir da reflexão aprofundada de conceitos advindos de outras áreas do conhecimento.

A ciência da informação busca soluções para responder à necessidade informacional da sociedade enquanto ciência social aplicada, tendo nesse contexto,

> a preocupação de esclarecer um problema social concreto, o da informação, e voltada para o ser social que procura informação, coloca-se no campo das ciências sociais (das ciências do homem e da sociedade), que são o meio principal de acesso a uma compreensão do social e do cultural. (LE COADIC, 1996, p. 21).

A ciência da informação refere-se à atividade direcionada à pesquisa de princípios e métodos que são partes da análise, do projeto e da evolução dos sistemas de informação. Nesses sistemas, os elementos constituintes são o ambiente, as pessoas, os recursos informacionais, as tecnologias e os procedimentos. Eles sustentam a capacidade para a busca de soluções e tomada de decisões como parte da vida diária, envolvendo a manipulação de dados, o acesso à informação e a apropriação do conhecimento.

A área é focada nas metodologias e nos instrumentos desenvolvidos ao longo do tempo para armazenar, descrever, recuperar, preservar, disseminar e compartilhar as experiências humanas.

Nesse contexto, a pretensão deste estudo é apresentar conceitos sobre dado e granularidade na ciência da informação sob o enfoque das tecnologias de informação e comunicação (TIC) e sua natureza interdisciplinar, trazendo à luz elementos que contribuam para uma reflexão mais aprofundada sobre o perfil dessa área do conhecimento, a partir dos trabalhos desenvolvidos no interior do grupo de pesquisas dos autores.

A noção de conceito é bastante importante para o entendimento da formalização teórica no universo científico, considerando que um constructo científico é sempre vinculado a uma fase filosófica (KUHN, 2005).

Na literatura da ciência da informação, entretanto, há uma pluralidade conceitual sobre os vocábulos que neste artigo se pretende tratar: dado e granularidade, particularmente na perspectiva de investigação e de trabalho com a temática informação e tecnologia.

DADO

A análise do conceito de dado permite refletir sobre um elemento amplamente utilizado e que tem tido sua definição relegada a segundo plano ou sendo conceituado de acordo com o contexto no qual o vocábulo é adotado.

O conceito de dado precisa ser redimensionado e se faz necessário que seja entendido e percebido como elemento básico nos fluxos informacionais, especialmente em um momento em que tanto se discute e se legisla sobre o acesso a dados na administração de conteúdos, no favorecimento de sua visibilidade e na sua utilização e reutilização.

Muitas foram as tentativas de definir o conceito de dado, e uma das dificuldades inerentes à definição reside na condição do termo e do próprio dado serem utilizados de maneiras distintas por praticamente todas as áreas da ciência. Neste texto, busca-se uma conceituação que atenda às necessidades da ciência da informação, sem deixar de considerar sua interface com a ciência da computação e com a comunicação.

Dentre as definições de dados, destaca-se a forte vinculação com o registro sobre fatos, tal como propõem Blumenthal (1969) e Fry & Sibley (1976), quando definem dados como um conjunto de fatos. Outra vinculação presente nas definições é com o meio de obtenção dos dados, conforme sugerem Davis e Rush (1979), "a forma mais simples de definir dados é dizendo que eles são o resultado de mensuração ou observação" ou, ainda, quando Yovits (1981) aponta que "dados são fatos ou podem ser ditos e acreditados como fatos que resultam da observação de fenômenos físicos".

A vinculação do conceito de dados com o processo de geração de informação, que apresenta maior carga semântica, é muito presente, principalmente na ciência da computação, ao considerar a informação como elemento intermediário entre o dado (básico e estruturado) e o conhecimento (complexo e com alta carga semântica), conforme reforça a definição de Dorn (1981), de que os dados são a matéria-prima para o desenvolvimento de informações. No entanto, a distinção entre os conceitos de dado e informação, nesta perspectiva, pode trazer circularidade na definição.

Outro aspecto relevante na definição de dado passa pela aceitação da característica simbólica. Segundo Burch et al. (1983), dados são substitutos simbólicos, linguísticos, matemáticos ou outros, que são aceitos como atributos para representação de pessoas, objetos, eventos e conceitos. Um indício interessante nesta concepção é a aceitação da característica representacional dos dados, apesar da implícita limitação que a questão do símbolo traz ao conceito, já que uma imagem ou determinado som também poderiam ser considerados dados.

Le Coadic (1996) trata do tema com a definição da informática, em que dado é a representação convencional codificada de uma informação, sob uma forma que permite seu processamento eletrônico, o que abre bastante a ideia de dado como representação e com foco no formato que o suporte digital exige, mas não apresenta a delimitação sobre

o que seria dado, o que conduz a possibilidade de se identificar qualquer conjunto de *bits* registrados, por maior que seja, como um dado.

Levy (2011), tomando por base o modelo cognitivo, opina que o dado não é definido por seu conteúdo, mas por seu sistema de endereçamento, concluindo que será considerado como tal quando for endereçado por uma URL, que simplesmente provê uma rota de acesso para um conteúdo único.

Davenport (1998) retoma a ideia a partir da definição de que dados são simples observações sobre o estado do mundo, podendo ser facilmente estruturados, obtidos, tratados e transferidos por máquinas. Ainda mais utilitarista, e baseado na visão da área de tecnologia, o autor retoma o conceito, descrevendo dados como "conjunto de fatos distintos e objetivos, relativos a eventos, sendo descritos utilitariamente, no contexto organizacional, como registros estruturados de transações" (DAVENPORT; PRUSAK, 1999, p.2).

Complementando, com destaque à questão da carga semântica dos dados, temos a definição de que "dados representam observações ou fatos fora de contexto e, portanto, sem uma significação direta" (ZACK, 1999).

Sob o ponto de vista do contexto de desenvolvimento de soluções para gerenciamento de bases de dados, identifica-se, como um marco, a definição do modelo relacional, que em sua gênese teve como um dos objetivos a independência dos dados, ou como indicou Codd (1981), "a motivação mais importante deste trabalho de pesquisa [proposta do modelo relacional] é proporcionar uma distinção clara entre aspectos físicos e lógicos da gestão de dados".

Santos e Santana (2002) conceituaram o termo dado "como um elemento básico, formado por signo ou conjunto finito de signos que não contém, intrinsecamente, um componente semântico, mas somente elementos sintáticos", que incorporava a questão do dado como símbolo, e como elemento que não contém semântica suficiente para ser interpretado por si só, minimizando, naquele momento, o problema da delimitação.

Entretanto, ao analisar dado como elemento que pode ser diretamente tratado por instrumentos digitais, considera-se sua composição baseada na tríade entidade - atributo - valor <e,a,v> (BOOCH; RUMBAUGH; JACOBSON, 1999), ou seja, cada conjunto mínimo de símbolos que pode ser tomado como uma unidade de conteúdo, precisa ser identificado com o contexto a que pertence.

A tríade entidade - atributo - valor (EAV) é base para modelagem de dados e bastante utilizada no mapeamento de dados heterogêneos, principalmente em aplicações de dados da saúde (NADKARNI et al., 1999). Citada nos textos de inteligência artificial (WINSTON, 1992), teve sua origem no conceito de listas de associação utilizadas na linguagem LISP.

Denominada *alist*, compreende uma estrutura de dados que registra uma associação entre um valor e uma chave. A associação (cons) é construída a partir de dois parâmetros (chave . valor), que na linguagem LISP recebem a denominação 'CAR' e 'CDR' remetendo as origens do uso da linguagem quando se utilizava comandos os mnemônicos assembler "Conteúdo do Registrador de Endereço" (*Contents of the Address part of Register number*) e "Conteúdo do Registrador de Decremento" (*Contents of the Decrement part of Register number*) respectivamente, compondo assim o registro de uma relação por meio destes dois parâmetros (car cdr) (SEBESTA, 2011, p. 697).

Tal prática favoreceu a implementação da informação sobre a entidade/objeto, a partir da utilização do conceito de associação e de relacionamento na estruturação de dados, propiciando que os conteúdos fossem persistidos em conjuntos. A utilização de estruturas complementares para a descrição das entidades e atributos permite a síntese da representação do contendo a partir da identificação da descrição da entidade e a identificação da descrição do atributo, vinculando-os ao terceiro elemento que registra o conteúdo (valor) (EAV).

Dados clínicos coletados a partir dos dados disponíveis no NICTIZ (2014), organização holandesa, especialista na elaboração e na gestão de padrões para intercâmbio de dados eletrônicos da saúde, são apresentados a seguir a título de ilustração.

Os dados são demonstrados por meio de relacionamentos com as tabelas descritivas que identificam as entidades (tabela 1) e os atributos (tabela 2), que referenciados pelas respectivas colunas são contextualizados e apresentados na figura 3.

A tabela 1 descreve as entidades que fazem parte do contexto sobre o qual os dados serão estruturados, contendo um código identificador único, o nome de cada entidade. Neste caso, o código da entidade da qual cada entidade é parte, explicitando, portanto, as relações de todo-parte existente entre as entidades.

Na tabela 2, são identificados os atributos que podem fazer parte do contexto dos dados, com um código único para cada atributo.

Tabela 1 - Tabela de entidades

Entidade		
Id	**Nome**	**Parte_De**
1	ClinicalDocument	
2	structureBody	1
3	section	2
4	entry	3
5	organizer	4
6	componente	5
7	observation	6
8	statusCode	7
9	effectiveTime	7
10	value	7
11	code	7
12	tempateId	7
13	text	7
14	reference	13
15	interpretationCode	7
16	id	7

Fonte: Adaptado de Lenz; Beyer e Kuhn (2007)

Tabela 2 - Tabela de atributos

Atributo	
Id	**Nome**
1	code
5	typeCode
6	classCode
7	moodCode
8	value
9	xsi.type
10	codeSystem
11	codeSystemName
12	unit
13	displayName
14	root

Fonte: Adaptado de Lenz; Beyer e Kuhn (2007)

Tabela 3 - Tabela de valores

Valores				
Id	**Fonte**	**Entidade**	**Atributo**	**Valor**
10	1	4	5	DRIV
11	1	7	6	OBS
12	1	7	7	EVN
13	1	8	1	completed
14	1	9	8	20000407
15	1	10	9	PQ
16	1	10	8	145
17	1	10	12	mm[Hg]
18	1	11	1	9999-9
19	1	11	10	9.99.999.9.9999999.9.9
20	1	11	11	LOINC
21	1	11	13	Intravascular Systolic
22	1	12	14	2.16.840.9.9999 99.99.99.99.9.99
22	1	16	14	c6f99999-99ad-99db-bd99
23	1	14	8	#vit6
24	1	15	1	N
25	1	15	10	9.99.999.9.999999.9.99

Fonte: Adaptado de Lenz; Beyer e Kuhn (2007)

A figura 1, por sua vez, expressa os valores e por conseguinte os dados, que podem ser identificados por um código, relacionando a cada valor o código identificador da entidade e o código identificador do atributo, que desse modo, contextualizam o conteúdo (valor). Nesta tabela são apresentadas, ainda, informações adicionais como fonte dos dados, cabendo, se necessário, a inclusão de outros metadados.

Os dados exibidos na tabela 3, ao serem instanciados, seguindo um padrão como o Clinical Document Architecture CDA, podem ser representados em eXtensible Markup Language (XML) e seguindo uma versão, como HL7, podem resultar na instância exemplificada na figura 1.

Figura 1 - Exemplo de registro de dados em HL7 CDA

```
<ClinicalDocument>
  <structuredBody>
    <section>
      <entry typeCode="DRIV">
        <organizer>
          <component>
            <observation classCode="OBS" moodCode="EVN">
              <statusCode code="completed"/>
              <effectiveTime value="20000407"/>
              <value xsi:type="PQ" value="145" unit="mm[Hg]"/>
              <code code="9999-9" codeSystem="9.99.999.9.999999.9.9" codeSystemName="LOINC" displayName="Intravascular Systolic"/>
              <templateId root="2.16.840.9.999999.99.99.99.9.99"/>
              <id root="c6f99999-99ad-99db-bd99"/>
              <text>
                <reference value="#vit6"/>
              </text>
              <interpretationCode code="N" codeSystem="9.99.999.9.999999.9.99"/>
            </observation>
          </component>
        </organizer>
      </entry>
    </section>
  </structuredBody>
</ClinicalDocument>
```

Fonte: Adaptado de Lenz; Beyer e Kuhn (2007) e Nictiz (2014)

No contexto da modelagem e da estruturação de dados, encontramos subsídios para mostrar um conceito de dados que considere a questão da interpretação semântica em sua constituição mais elementar.

Assim, se considerarmos como exemplo o conjunto de símbolos '204', por si só, não teria como ser tratado por um algoritmo. No entanto, se for identificado como sendo o número de páginas de determinado livro, teremos a tríade: <livro, número de páginas, 204>, ou <e,a,v>, ou seja, para determinado dado poderia ser identificada uma carga semântica inicial como um valor, a quantidade 204, vinculado ao atributo número de páginas de uma entidade livro.

Há de se considerar, ainda, outros elementos que surgem dessa definição, como, por exemplo, o conjunto de valores válidos para determinado atributo que define o domínio deste atributo. Elementos como esses configuram-se como uma contextualização de determinado dado, e, nesse ponto de vista, o dado não é totalmente desprovido de semântica.

Destaca-se também a necessidade de ampliar ou explicitar a designação do aspecto simbólico da composição de um dado, mantendo, ainda, a delimitação de dados no contexto da ciência da computação. Conforme destaca Hughes (2009), dados podem representar textos, por meio de cada

caractere, representado por um código numérico único, ou podem representar imagens pela codificação de cada pixel por meio de codificação numérica para sua intensidade, cor, ou combinação das duas, concluindo que, por meio de uma codificação numérica apropriada, um computador pode processar qualquer tipo de dado. Não importa o que esteja sendo representado pelos dados, para o computador, serão sempre números.

Com relação ao caráter representacional dos dados, destaca-se a distinção entre o esquema <e,a> e a instância <v> que compõem a tríade <e,a,v> de um dado.

A estrutura semântica básica e as regras sintáticas presentes estão representadas no esquema <e,a> de determinado dado e permitem que eles sejam interpretados e utilizados em determinado contexto. Para que os dados sejam corretamente utilizados, o acesso ao esquema correspondente deve ser de domínio do utilizador, e eles podem estar definidos de forma tácita ou explícita, ou seja, apenas como parte do conhecimento prévio do utilizador (tácito) ou disponível sob a forma de metadados armazenados (explícito) para que o dado seja obtido e tratado quando necessário. A importância do esquema é notada em um contexto em que a interoperabilidade deve ser um dos pré-requisitos para a disponibilização de dados.

Instância <v> por sua vez identifica a aplicação da tríade definidora de dados <e,a,v> em um caso concreto, em que não só a definição da entidade <e> e do atributo <a> são conhecidas, mas também o valor específico do atributo de um elemento da entidade.

Dados, como parte da informação, são compostos pela tríade <e,a,v>, estruturados por um esquema <e,a> que os contextualiza, tácita ou explicitamente, e são identificados como dados por representarem a granularidade mais fina possível de determinado contexto de uso, quando a tríade se completa com o valor <v> no momento de sua instanciação.

Como ilustração, pode-se pensar em um tipo documental de entidade livro<e> que contém o atributo título<a> que se instancia pelo valor "As Tecnologias da Inteligência"<v> como um elemento da entidade de um caso concreto. Ou ainda, exemplificando com a inclusão de outros atributos desta entidade, podemos considerar o caso deste livro<e> que tem como autor<a> "Pierre Lévy"<v>, com total de páginas<a> 204<v>, do qual se pode abstrair a seguinte estrutura:

<e,a,v>
<Livro,Título,As Tecnologias da Inteligência>
<Livro,Autor,Pierre Lévy>
<Livro,Página,204>

Destacando que, para ser utilizada, tal estrutura precisará ter sua unidade garantida por vinculação à entidade e ao(s) atributo(s) que a identificam e qualificam, e que pode ser representada por meio de uma linguagem como a *Extensible Markup Language* - XML:

```
<documento>
<título>As Tecnologias da Inteligência</título>
<autor>Pierre Lévy</autor>
</documento>
```

Conclui-se que dado é uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação.

Esta conclusão encaminha a uma reflexão sobre dados estruturados e não estruturados.

DADOS ESTRUTURADOS E NÃO ESTRUTURADOS

Ao adotarmos a definição de estrutura proposta por Eco (2007), como um conjunto de elementos que propicia o estabelecimento de uma relação de denotação fixada pelo código, e levando-se

em conta que tal estrutura pode ser representada pela relação com metadados ou mesmo por meio da vinculação explícita ou implícita com aspectos como sua semântica posicional, pode-se afirmar que é por meio da estrutura que se pode viabilizar a utilização e a reutilização de dados.

Assumindo que a estrutura de um dado é estabelecida pela identificação dos seguintes elementos <e> e <a>, que pertencem à tríade que compõe o seu esquema, a questão que surge é a diferenciação entre dados estruturados e não estruturados.

Ao analisar dois exemplos de dados, tais como um livro e um registro bibliográfico, ambos em formato digital, tem-se a percepção de que o livro é não estruturado e que o registro bibliográfico é estruturado. O que configura a diferenciação é que, no registro bibliográfico, sua estrutura interna está explícita, ou seja, seu conteúdo é fixado por um código, e no livro a identificação de sua estrutura requer um interpretador externo, o que o configura como um dado não estruturado. O mesmo poderia ser aplicado a uma música ou vídeo disponíveis em ambientes digitais.

Ao acessar um arquivo em formato AVI, para saber se seu conteúdo <v> é uma música, uma entrevista ou um filme, é necessária a ação de um interpretador externo. Nesse momento, pode-se destacar que, mesmo possuindo alguns elementos de estrutura embebidos em seu conteúdo, tais como metadados de origem e de tecnologia utilizada na compactação, essa estrutura não é relevante semanticamente para a interpretação desejada com relação ao seu conteúdo.

Já um dado estruturado como registro bibliográfico tem seu conteúdo interno explicitado por meio de uma semântica que permite que todo seu conteúdo seja interpretado de forma autônoma por máquina. O mesmo pode ser aplicado a um arquivo de intercâmbio de informações entre instituições, por exemplo, os que são transmitidos sobre informações de cobrança entre o cedente e a instituição bancária, pois o que contextualiza cada item transmitido é a sua posição relativa dentro de cada linha, ou seja, a estrutura é representada por meio de semântica posicional. Já no caso de um arquivo de *Really Simple Syndication* (RSS) de determinado sítio, temos cada um dos dados que o compõem identificados por marcadores estruturados em XML, o que permite, inclusive, a visualização de seu conteúdo por interfaces genéricas e preparadas para tal fim.

Nesse contexto, a interpretação de dados, especialmente de dados não estruturados, requer a utilização de dados sobre os dados, ou seja, de metadados.

A representação dos dados em um sistema de informação se dá pela utilização de metadados que, de acordo com Alves (2010, p. 47-48):

> [...] são atributos que representam uma entidade (objeto do mundo real) em um sistema de informação. Em outras palavras, são elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades; são ainda dados que descrevem outros dados em um sistema de informação, com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação.

A relevância dos metadados é percebida na complementação da estrutura semântica mínima, da tríade <e,a,v> que compõe determinado dado, complementação que revela, inclusive, o nível de detalhe presente no dado e que pode ser entendido por uma análise de sua granularidade.

GRANULARIDADE

Na proposta de analisar o conceito de dados e suas implicações para a ciência da informação estão os aspectos relacionados aos conjuntos de dados, como a granularidade, a cardinalidade e a diversidade (NEHRING; PUPPE, 2002).

Todo conjunto de dados que tenha cardinalidade maior que 1 (#S > 1), pode ser analisado e tratado a partir de possíveis subconjuntos que possam ser identificados e abstraídos a partir de determinado critério de dissimilaridade com uma ou mais dimensões de análise. Assim, a granularidade de um conjunto de dados está vinculada ao número de atributos que o compõem e a diversidade de seus conteúdos.

Quanto maiores as possibilidades de obter subconjuntos a partir de um conjunto de dados, maiores serão as potencialidades de tratamento e de elaboração de resultados; segundo Nehring & Puppe (2002), a granularidade (Γs) de um conjunto S pode ser entendida como a mínima dissimilaridade entre dois objetos distintos de S:

$$\Gamma s := \min_{x,y \in S,\, x <> y} d(x,y)$$

e, quanto maiores as possibilidades, maior será o detalhamento disponível, o que determina a definição da granularidade como "fina". No caso de menor detalhamento, tem-se a granularidade "grossa" (WANG; WU, 2003).

Por exemplo: se um conjunto de dados estruturado (T) contiver apenas as médias finais de cada turma de alunos de uma escola, a granularidade será mais grossa que outro conjunto de dados (D) que contenha as médias de notas por disciplina de cada turma. E a granularidade pode ser ainda mais fina em um conjunto (A), se estiverem disponíveis as médias por aluno.

Ou, ainda, se um conjunto de dados estruturado (C) contiver apenas as médias finais de empréstimos de livros de cada biblioteca de uma cidade, a granularidade será mais grossa que outro conjunto de dados (U) que contenha as médias de empréstimos de livros por tipos de usuários de cada uma das bibliotecas. E a granularidade pode ser ainda mais fina em um conjunto (B), se estiverem disponíveis as médias de empréstimos de livros por usuário de cada biblioteca.

Assim, a inclusão de novos atributos aos objetos que compõem o conjunto pode propiciar a identificação de dissimilaridades e, no exemplo apresentado, o conjunto (T), com os atributos turma e média, tem menor detalhamento que o conjunto (D), que teria de ser composto, pelo menos, pelos atributos turma, disciplina e média. Por sua vez, o conjunto (D) teria menor detalhamento que o conjunto (A), que teria os atributos turma, disciplina, aluno e média.

Destaca-se que o ganho em detalhamento tem, como contrapartida, o aumento do custo de armazenamento e de processamento (tratamento) em função do aumento da cardinalidade desses conjuntos (KIMBALL; STREHLO, 1995).

A granularidade está vinculada ao conteúdo disponível no conjunto de dados e impacta diretamente nos processos de acesso e de tratamento. Já a visualização poderá ou não fazer uso desse potencial, tornando-o acessível ou não. No exemplo mostrado, uma consulta que apresente as médias por turma terá os mesmos resultados se construída sobre o conjunto (T) ou sobre o conjunto (A). No entanto, se o usuário desejar conhecer quais as médias por disciplina ou por aluno de cada uma das turmas, só terá esse resultado se a visualização for feita sobre o conjunto com granularidade mais fina (conjunto A) e se essa funcionalidade (tratamento) estiver disponível.

Conjuntos de dados com granularidade fina podem ser subutilizados se as visualizações não contiverem funcionalidades que permitam a consulta aos subconjuntos. No entanto, conjuntos de dados com granularidade grossa diminuem as possibilidades de consulta e podem reduzir a relevância de recursos de visualização. Assim, no acesso e na visualização de conjuntos de dados, é necessário que o usuário tenha conhecimento sobre as funcionalidades disponíveis no recurso de visualização e sobre as potencialidades de utilização do conjunto de dados utilizado, e esse é um aspecto de impacto direto na ciência da informação, seja na arquitetura da informação, seja na formação de profissionais que irão trabalhar no projeto e na utilização de recursos de acesso a dados.

Destaca-se também que a semântica da definição do atributo e, por conseguinte, de uma dimensão passível de análise, não define a granularidade de um conjunto de dados, ou seja, se determinado atributo define, por exemplo, o tamanho de um objeto, o fato de a dimensão apresentar o conteúdo em centímetros quadrados não indica que a granularidade é mais fina do que se o conteúdo fosse apresentado em metros quadrados, já que

o que muda é somente a unidade de medida, diferente do caso de atributos que apresentem outras informações sobre a dimensão, como altura e largura, o que permitiria a identificação de subconjuntos menores, potencializando, como consequência, o acesso a detalhes da informação, tendo, neste caso, granularidade mais fina.

Na ciência da informação, os metadados têm papel preponderante, já que é por meio de dados sobre os dados que se pode representar, de modo a descrever e a identificar conteúdos (SANTOS; ALVES, 2009, 2013). Analisando conjuntos de metadados, a definição do menor subconjunto passível de ser identificado define a granularidade dos metadados utilizados.

No processo de recuperação da informação, um dos temas centrais da ciência da informação, a granularidade tem papel de destaque no estudo sobre interfaces para resultados de buscas em sistemas de recuperação orientados a consultas hierárquicas, auxiliando o entendimento e a exploração do contexto da informação recuperada, propiciando, por exemplo, o destaque da posição na hierarquia granular e expondo a relação com similares na hierarquia (BALATSOUKAS; DEMIAN, 2010).

CONSIDERAÇÕES FINAIS

Com o aumento da demanda por acesso a dados no desenvolvimento de pesquisas e de novas tecnologias, o objetivo deste artigo é suscitar a reflexão e o debate sobre o conceito de dados, de conjunto de dados e de granularidade no domínio da ciência da informação, que se caracteriza como um ambiente de pesquisa baseado na geração e no uso intensivo de dados, desencadeado principalmente pelo desenvolvimento de instrumentos científicos e pela aplicação de instrumentos tecnológicos, como computadores.

O conceito de dados pode gerar avanços significativos para a área de ciência da informação, pois representa um problema de grande impacto e abrangência para a atividade de pesquisa, reforçando os pressupostos da ciência da informação. Nessa perspectiva, o tratamento da informação, a representação de recursos, a recuperação e a disseminação de informação se tornam áreas vinculadas à descrição, ao armazenamento, à preservação, ao acesso e à gestão de dados.

Os conceitos aqui apresentados podem ser considerados uma novidade em termos de aplicação e de estudos no âmbito da ciência da informação no país. Porém, neste trabalho, pretendeu-se apenas apresentar o conceito de dados, esclarecendo suas características e estrutura.

Por fim, é importante ressaltar que estudos iniciais já apontam para alguns desafios que merecem ser investigados academicamente com mais profundidade e de modo interdisciplinar: mecanismos de recuperação; apresentação e navegação dos dados e suas ligações, por meio de mecanismos visuais que usem taxonomia ou ferramentas de visualização semântica; confidencialidade e acesso aos dados; acesso livre ou não aos dados; formatos de metadados para a gestão da preservação do significado e da estrutura de coleções de dados de pesquisa em ambientes de repositórios de dados; recuperação dos dados e processamento de conhecimento de forma inteligente; ciclo de vida dos dados; compartilhamento de dados, entre outros.

REFERÊNCIAS

ALVES, R.C.V. ; SANTOS, P.L.V.A.C. *Metadados no domínio bibliográfico*. Rio de Janeiro: Intertexto, 2013. 196p.

ALVES, R.C.V. *Metadados como elementos do processo de catalogação*. 2010 (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

BALATSOUKAS, P. ; DEMIAN, P. Effects of Granularity of Search Results on the Relevance Judgment Behavior of Engineers: Building Systems for Retrieval and Understanding of Context. *Journal of the American Society for Information Science and Technology*, v.61, n.3, p.453–467, 2010.

BLUMENTHAL, S. C. *Management information systems*. Englewood Cliffs: Prentice-Hall. 1969.

BOOCH, G.; RUMBAUGH, J.; JACOBSON, I. *The unified modeling language user guide*. Reading, Mass.: Addison-Wesley, 1999.

BURCH, J.G.; STRATER, F.R.; GRUDNITSKI, G. *Information Systems*: theory and practice. 3. ed. New York: John Wiley & Sons,1983.

CODD, E.F. *Data models in database management*. ACM SIGPLAN Notices. 1981.

DAVENPORT, T. *Ecologia da informação*: porque só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998.

DAVENPORT, T.; PRUSAK, L. *Conhecimento empresarial*. Rio de Janeiro: Campus, 1999.

DAVIS, C. H.; RUSH, J. E. *Guide to information science*. Westport: Greenwood Press.1979.

DORN, P.H. Business information in the eighties. In: PAPPENHEIM, A.E.(Ed.) *Business information systems*, . Maidenhead: Pergamon Infotech,1981. p. 245-260. (INFOTECH State of the Art Report, series 9, n.7)

ECO, Humberto. *A estrutura ausente*: introdução à pesquisa semiológica. São Paulo: Perspectiva, 2007.

FRY, J.P.; SIBLEY, E.H. Evolution of data-base management systems. *ACM Compufing Surveys*, v.8, n.1, p. 7-42, 1976.

HUGHES, J. M. Embedded Image Data Procesing on Mars. In: SEGARAN, T.; HAMMERBACHER, J. *Beautiful Data*: the stories behind elegant data solutions. Sebastopol: O'Reilly, 2009.

KIMBALL, R.; STREHLO, K. Why decision support fails and how to fix it. *ACM SIGMOD Record Homepage archive*, v.24, n.3. New York: ACM, sept. 1995. p. 92-97. DOI:10.1145/211990.212023.

KUHN, T. S. *A estrutura das revoluções cientificas*. 9. ed. São Paulo: Editora Perspectiva. 2005. (Coleção Debates Ciência)

LE COADIC,Y.F. *A Ciência da Informação*. Brasília, D.F: Briquet de Lemos. 1996.

LENZ R.; BEYER M.; KUHN K.A. Semantic integration in healthcare networks. *International Journal of Medical Informatics*, v.76, n.2-3, fev./mar., 2007. p. 201-207. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1386505606001171> Acesso em: 01 jun. 2014.

LEVY P. *The Semantic Sphere 1*: computation, cognition and information economy. Canada:Wiley-ISTE. 2011.

MORIN, E. *Liberdade e complexidade*. São Paulo: Associação Palas Athena, 1998, p.12-19. (Ensaios Thot, n.67)

NADKARNI, P.M.;MARENCO L.; CHEN R.; SKOUFOS E.; SHEPHERD G.;

NEHRING, Klaus; PUPPE, Clemens. A Theory of diversity. *Econometrica*, v.70, n.3, may., 2002. p.1155–1198.

NICTIZ. Good Health Health Summary. 2014. Disponível em: <https://decor.nictiz.nl/CDA/_UV_generic_examples/Sample1.CDAr2.xml> Acesso em: 05 jun. 2014.

SANTOS, P.L.V.A.C.; ALVES, R.C.V. Metadados e web semântica para estruturação da web 2.0 e web 3.0. *DataGramaZero - Revista de Ciência da Informação*, v.10, n.6, 2009. Disponível em: <http://www.dgz.org.br/dez09/Art_04.htm>. Acesso em: 15 jun. 2012.

SANTOS, P.L.V.A.C.; SANT'ANA, R.C.G. Transferência da informação: análise para valoração de unidades de conhecimento. *DataGramaZero -Revista de Ciência da Informação*,v.3, n.2, 2002. Disponível em: <http://www.dgz.org.br/abr02/Art_02.htm>. Acesso em: 15 jun. 2012.

SARACEVIC, T. Interdisciplinary nature of information science. *Ciência da Informação*, Brasília, v.24, n.1, p.36-41, 1995.

SEBESTA, R.W. *Conceitos de linguagens de programação*. Bookman: São Paulo. 2011.

WANG, L. ; WU, G. Attribute reduction and information granularity. *Journal of Systemics, Cybernetics and Informatics*. International Institute of Informatics and Cybernetics, v.1 n.1, p. 32-37. 2003

WERSIG, G. Information science: the study of postmodern knowledge usage. *Information Processing and Management*, v.29, n.2, p.229-239, 1993.

WINSTON P.H. Artificial intelligence. 3. ed. [s.l.]: Addison-Wesley, 1992.

YOVITS,M.C. Information and data. In: RALSTON, A.(Ed.) *Encyclopedia of computer science and engineering* . New York: Van Nostrand Reinhold, 1981. p. 714-717.

ZACK, M. Management codified knowledge. *Sloan Mangement Review*, v.40, n.4, summer, 1999.