DOI: 10.5433/1981-8920.2021v26n4p146

O PAPEL DOS VOCABULÁRIOS NO ACESSO E REUSO DOS *BIG DATA*

THE ROLE OF VOCABULARIES TO THE ACCESS AND REUSE OF BIG DATA

Carlos Henrique Marcondes[a]
Mauricio Augusto Cabral Ramos Junior[b]
Sergio de Castro Martins[c]

RESUMO
Objetivo: De forma similar à "explosão informacional" o fenômeno do *Big Data* vem sendo de forma crescente, objeto da CI/OC. Como descobrir, acessar, processar e reusar a enorme e crescente quantidade de dados que são disponibilizados continuamente na *Web* por nossa sociedade? Em especial, como tratar os chamados "dados não estruturados", documentos textuais, que sempre foram o objeto da CI/OC? **Metodologia**: Teorias de amplo espectro como Ontologia e Semiótica foram utilizadas para analisar dados como elemento essencial do *Big Data*, em especial os "dados não estruturados". **Resultados**: A partir da análise de várias definições de dados, um dado é identificado como parte de esquemas lógicos e semióticos já conhecidos, as proposições. Um dado é encontrado juntamente com outros, formando conjuntos de dados. Conjuntos de dados são na verdade conjuntos de proposições. Estas estão presentes no que é conhecido como dados estruturados - tabelas de bancos de dados relacionais ou de planilhas. Documentos textuais também contém conjuntos de proposições. Dados estruturados são comparados com "dados não estruturados". **Conclusões**: Embora no limite, ambos contenham proposições e possam ser equivalentes, enquanto conjuntos, dados estruturados são expressos e percebidos como um todo, conjuntos de dados não estruturados são processuais, expressos sequencialmente o que torna mais difícil a identificação de dados não estruturados em documentos textuais para seu processamento por máquinas.

Descritores: *Big Data*. Vocabulários. Dados estruturados. Dados não estruturados. Dados Abertos Interligados.

[a] Doutor em Ciência da Informação pela Universidade Federal do Rio de Janeiro (UFRJ). Docente Titular do Departamento de Ciência da Informação e do PPGCI da Universidade Federal Fluminense (UFF) e pesquisador 1D do Conselho Nacional de Desenvolvimento Científico e Tecnológico. E-mail: ch_marcondes@id.uff.br

[b] Doutorando em Ciência da Informação na Universidade Federal Fluminense (UFF). E-mail: macrjunior@gmail.com

[c] Doutor em Ciência da Informação pela Universidade Federal Fluminense (UFF). Mestrando em Filosofia pela Universidade do Estado do Rio de Janeiro (UERJ). E-mail: sergio.scm@gmail.com

1 INTRODUÇÃO

Já há algum tempo tem sido destacado, tanto na literatura comercial quanto na científica, a quantidade de dados produzidos pela sociedade atual, seu crescimento "explosivo", as dificuldades de armazená-los e reusá-los. Este fenômeno é apontado como "dilúvio de informações" (*information deluge*), "dilúvio de dados" (*data deluge*), "*tsunami of data*" (HEY; TREFETHEN, 2003), ou mesmo *Big Data*. Segundo estas mesmas fontes, o fenômeno tem impactado consideravelmente os negócios, governo e sociedade. Conforme apontado pelo *FAIR Compliant Biomedical Metadata Templates*,

> The ultimate Big Data challenge lies not in the data, but in the metadata — the machine-readable descriptions that provide data about the data. It is not enough to simply put data online; data are not usable until they can be 'explained' in a manner that both humans and computers can process." (FAIR - Compliant Biomedical Metadata Templates, 2019).

Além das questões como espaço de armazenamento e facilidade de acesso, o *Big Data* coloca uma questão mais geral, fundamental para o perfil da sociedade contemporânea. De acordo com a empresa *Cognizant*, "*While data volume proliferates, the knowledge it creates has not kept pace*" (COGNIZANT, 2011). Neste sentido, existe um grande potencial no *Big Data* que ainda não explorado: como reusar esses dados como recurso econômico, social, cultural, educacional, científico etc, para as mais diferentes atividades? De qualquer maneira, uma questão parece certa: não conseguiremos aproveitar todo o potencial do *Big Data* sem a utilização de meios técnicos, ferramentas e metodologias. Uma perspectiva nesta direção é a proposta do projeto da Web Semântica de agenciar máquinas para processarem o conteúdo da Web (BERNERS-LEE; HENDLER; LASSILA, 2001). Neste cenário, que contribuições a Ciência da Informação (CI) e a Organização do Conhecimento (OC) podem oferecer? A CI e a OC poderiam lidar com os dados do *Big Data*? Estão instrumentalizadas teórica e praticamente para enfrentar este desafio? Como este fenômeno se relaciona com informação e conhecimento?

O quadro atual nos remete naturalmente para a chamada "primeira explosão da informação", fenômeno fundador da CI, sobretudo com a

proliferação desordenada da literatura científica, em especial a partir dos anos 50-60 do século XX. Neste contexto, os Sistemas de Organização do Conhecimento (SOC) têm sido ferramentas fundamentais para tratar os registros de trabalhos científicos que começaram a ser armazenados em bases de dados computadorizadas. SOCs eram utilizados para atribuir assuntos de forma padronizada na entrada dos Sistemas de Recuperação de Informação (SRI) e na sua saída, na recuperação de informações propriamente dita, de maneira a auxiliar usuários na escolha de termos padronizados que representassem suas necessidades de informação a serem submetidas ao SRIs.

Na "primeira explosão da informação" e no lastro da experiência dos SOCs tradicionais para recuperação de informação temática em acervos de bases de dados e bibliotecas – como listas de termos, classificações biblioteconômicas enumerativas, classificações Facetadas etc - e das tradições teóricas, metodológicas e técnicas da Biblioteconomia, os SOCs tiveram que evoluir. Novas gerações de SOCs foram concebidas e construídas para lidar com a recuperação de informações em sistemas computacionais, como os tesauros - "*the information retrieval thesaurus*" (CLARKE, 2019). Há muito, desde fins do século XIX e início do século XX, o controle bibliográfico universal, o Instituto de Bibliografia de Bruxelas, o Repertório Bibliográfico Universal, a Classificação Decimal Universal e a própria área da Documentação, tal como concebida por Paul Otlet e La Fontaine, também foram respostas à problemática do excesso de meios de informação e da produção bibliográfica de uma cultura que se expandia e dependia cada vez mais da Ciência (RAYWARD, 1975).

Neste cenário os documentos e registros da informação tornaram-se objetos centrais da CI e OC, áreas cujo foco constituiu-se historicamente no problema da "explosão informacional" e da "recuperação da "informação". O pressuposto implícito na maneira como a CI e OC enfrentaram e enfrentam ainda hoje tais desafios é a recuperação da informação textual a partir de consultas e formulações de buscas que expressariam as "necessidades de informação" de um usuário, com grande potencial de atendimento desta demanda, ao serem processados e lidos pelo usuário. Atualmente, num entorno de informações com base na web, o problema se coloca em como atribuir a este conteúdo uma

semântica computacional (MARCONDES, 2012) para que as máquinas possam auxiliar no endereçamento na questão do *Big Data*.

A necessidade de busca por soluções computacionais que agregassem significados a conteúdos disponíveis na *Web* possibilitou a ampliação do conceito de SOCs para além dos vocabulários controlados temáticos e tesauros. Zeng (2019b) propõe a denominação geral de vocabulários e uma tipologia abrangente: 1) vocabulários de valores. Estes assinalam *valores padronizados* para um único campo ou propriedade do objeto sendo representado. Nos SOCs tradicionais este era os assuntos assinalados a objetos portadores de mensagens (CAPURRO, 2003), de temas ou de assuntos, como documentos[1]. Outros exemplos de vocabulários de valores seriam os vocabulários de autoridades - nomes padronizados de pessoas, famílias e instituições. O outro tipo de vocabulário proposto por Zeng (2019a) são os 2) vocabulários de metadados. Estes consistem num conjunto padronizado de *propriedades descritivas* do objeto a ser representado, como o conjunto de metadados *Dublin Core* (DC). Neste artigo, usaremos esta denominação abrangente de vocabulários.

Através da utilização das bases teóricas da Semiótica e Ontologia, este artigo tem por objetivo discutir o fenômeno do *Big Data*, além de propor sua análise e caracterização no escopo da CI e OC. Neste sentido, endereça as seguintes questões: O que é o *Big Data*? como caracterizá-lo e integrá-lo no escopo teórico da CI? Como organizar e dar sentido à grande e crescente quantidade de dados disponíveis na Web para permitir acesso, reuso e processamento automático? Qual o papel da CI/OC e, em especial, dos vocabulários nesta questão – uma das áreas de maior pesquisa teórica e aplicada da CI/OC?

A hipótese assumida aqui para responder a estas questões é que dados e metadados, constituindo como vocabulários, são um instrumento semântico e podem ter um papel fundamental para atribuir semântica computacional, de acordo com a proposta da Web Semântica, aos dados gerados pelo *Big Data*,

[1] Outros exemplos de vocabulários de valores seriam os vocabulários de autoridades - nomes padronizados de pessoas, famílias e instituições.

especialmente aos chamados dados não estruturados, como os dados textuais.

O artigo está organizado da seguinte maneira: após esta Introdução, a seção 2 expõe a Metodologia utilizada, a seção 3 analisa o fenômeno do *Big Data* a partir de exemplos concretos (descritos na seção 2), buscando uma definição de *dados*, relacionando dados com significado e com metadados. A seção 4 discute o papel dos vocabulários em dar semântica aos conteúdos disponíveis na Web, tanto para pessoas quanto para máquinas. A seção 5 faz as considerações finais e propõe futuras direções de pesquisa.

2 METODOLOGIA

Esta pesquisa usa metodologia dedutiva, utilizando bases teóricas, conceituais e metodológicas da Ontologia/Metafísica, da Semiótica Peirceana e da Lógica, para analisar os conceitos de *Big Data* e de dados. Os conceitos usados provêm das seguintes disciplinas: 1) Ontologia: Categorias (ARISTÓTELES, 2000), Entidades, Atributos, Relacionamentos (CHEN, 1976), (FURNER, 2014), Identidade/Essência, Independência existencial X Dependência (GUARINO, 1998); 2) Teoria do Conceito, Teoria das Definições, "campo conceitual" (DAHLBERG, 1978, 1981, 1992); 3) Lógica (MACIEL, 1974); 4) Semiótica Peirceana (PEIRCE, 1931), (SANTAELLA, 2008). Quanto aos procedimentos metodológicos, trata-se de uma pesquisa bibliográfica. Foram usados como objetos e referência para análise dois exemplos reais de conjuntos de dados, a saber, dados epidemiológicos de Covid-19 do Hospital das Clínicas da USP obtidos do repositório COVID-19 Data Sharing/BR mantido pela FAPESP[2], e dados meteorológicos do bairro do Jardim Botânico, cidade do Rio de Janeiro, do ano de 2020, obtidos do sistema Alerta Rio[3].

Definições de *Big Data* de artigos com origem na CI/OC foram usadas como orientação para análise dos exemplos reais. O *Google Acadêmico* foi utilizado como ferramenta de busca e o programa gestor de referências *Publish*

[2] Disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/100

[3] Disponível em: http://www.sistema-alerta-rio.com.br/dados-meteorologicos/download/dados-meteorologicos/

or Perish[4], usando como estratégia booleana os seguintes parâmetros: “big data” [title], “information Science”, “knowledge organization”, definition "big data is", “from 2010 to 2021”, “no citations”, “no patents”. De acordo com esta estratégia, foram recuperados 17 artigos, dos quais 10 foram descartados por não trazerem definições de *Big Data* ou não serem oriundos da área de CI/OC, resultando em 7 artigos, conforme consta do Anexo 1. Tais artigos basilares da área de CI/OC, trazendo definições de dados, um tema recorrente na área (ROWLEY, 2007), (HJØRLAND, 2018), (FLORIDI, 2019).

3 O FENÔMENO *BIG DATA*

Uma busca utilizando a estratégia de busca citada na seção 2 recuperou poucos artigos realmente oriundos da área de CI/OC que propusessem definições de *Big Data* (ver Anexo 1). Dos artigos com definições, poucos vão além de caracterizar *Big Data* como um fenômeno que envolve grandes quantidades de dados, heterogeneidade dos dados, fluxo contínuo de geração e atualização, necessidade de grande capacidade de processamento para revelação de padrões ou tendências. Estas definições amplas do fenômeno não são necessariamente prevenientes da CI/OC. Por este motivo, faz-se necessária uma definição oriunda das áreas de CI/OC visto que, se tais áreas objetivam desempenhar um papel nos problemas do entorno *Big Data*, um trabalho de conceitualização se faz necessário.

Qualquer definição ou caracterização de *Big Data* teria necessariamente que passar, mencionar ou qualificar também *Dados*. A essência do fenômeno *Big Data* são os dados, uma vez que não existe *Big Data* sem dados. Embora se constate a necessidade de um trabalho de conceitualização do *Big Data* pela área, com relação a *dados* a área revela, desde há algum tempo, um acúmulo conceitual. Exemplos disso são as várias definições de informação que a relacionam com dados, além das discussões acerca da famosa hierarquia *Dados-Informação-Conhecimento-Sabedoria* (ROWLEY, 2007), ou discussões

[4] Disponível em: https://harzing.com/resources/publish-or-perish.

mais recentes sobre curadoria dos dados de pesquisa[5] e o papel nesta questão das bibliotecas. Uma possível definição conceitual de *Big Data* certamente mencionaria também dados e provavelmente teria esta forma: "*Big Data: df* Dados que...". Estas questões possibilitaram uma abordagem de *Big Data* a partir dos dados (HJØRLAND, 2018).

3.1 Considerações sobre o Conceito de Dados

A terminologia em torno do termo *Dado* tem sido objeto de estudo de muitos autores, proporcionando um debate sob perspectivas diferentes e controversas. Para Hjørland (2018), há uma questão de natureza epistemológica relacionada à objetividade e subjetividade, ou seja, enquanto insumo para conhecimento, os dados não são neutros em si mesmos, o que pressupõe uma intencionalidade na sua captura ou produção. Foram muitas as tentativas de conceituar este termo, sobretudo pelo fato de definições serem explicitadas por várias áreas do conhecimento. No entanto, destaca-se aquelas relativas ao registro e recuperação de fatos, ao processo de geração da informação (seu caráter semântico), às características simbólicas e, também, aquelas oriundas da Informática.

Neste sentido, *Dado* é uma unidade de conteúdo obrigatoriamente referenciada a uma entidade e a seus atributos, sendo também o valor deste atributo. Entidade, atributo e valor são constituídos para representar um certo contexto, de tal maneira que seu conteúdo possa ser compartilhado e plenamente interpretado (SANTOS; SANT'ANA, 2013). Segundo este entendimento, dados são um fenômeno artificial, uma criação humana, um artefato, o que pressupõe uma intencionalidade (HØRLAND, 2018).

No que concerne a uma definição de dados, a conclusão de Hjørland (2018): "*Data is [are] information on properties of units of analysis*" (HJØRLAND, 2018) pode-se relacionar propriedades com predicados/proposições. Esta interpretação encontra compatibilidade com as definições da *Enciclopedia Stanford*, na qual: "*Properties are those entities that can be predicated of things*

[5] Mais detalhes em https://www.dcc.ac.uk/about/digital-curation, https://www.go-fair.org/

or, in other words, attributed to them. Thus, properties are often called predicables" (ORILIA, PAOLETTI, 2020). Hjørland (2018) analisa especificamente o fenômeno do *Big Data* e, ao fazê-lo, preocupa-se em analisar várias definições de dados, relacionando dados a um fenômeno (dados de um fenômeno), a observação deste, a intencionalidade desta observação, suas propriedades (do fenômeno), documentos como fonte de dados etc. Nestas considerações, o conceito de *dados* (plural, "data") é distinguido do de *dado* (no singular, "datum", o que seria uma unidade de dados), assim como os conceitos de *metadado x dado* - uma propriedade e seu valor específico para determinado fenômeno que está sendo analisado/observado. Recorre para isto a conceitos como fenômeno e propriedades retirados da Metafísica, área que dá suporte às outras ciências, conforme atesta Bunge (2015), a exemplo do que fazem as outras ciências ao analisarem seus objetos.

Como mencionado anteriormente, são bastante recorrentes na CI e OC colocações que vinculam dados, informação e conhecimento (ROWLEY, 2007). Estas entidades são colocadas como formando uma escalada de complexidade, ou, de alguma maneira, em uma visão processual, como se dados fossem subsídio para informação e esta, para conhecimento. Rowley (2007) fala de "hierarquia da sabedoria" e indica que o conceito de dados como símbolos tem origem em Ackoff:

> Data are defined as symbols that represent properties of objects, events and their environment. They are the products of *observation*. But are of no use until they are in a useable (i.e. relevant) form. The difference between data and information is functional, not structural (ACKOFF *apud* ROWLEY, 2007, p. 165).

Nesta definição em que são mencionados "representação de propriedades de objetos, eventos e seu ambiente", chamam atenção os aspectos Semiótico – representar - e Ontológico – propriedades.

Na Semiótica Peirceana, representar é "*To stand for, that is, to be in such a relation to another that for certain purposes it is treated by some mind as if it were that other*" (PEIRCE, 1931, CP 2.273). Isto pode ser entendido como uma relação dita triádica (entre três) envolvendo um *objeto* (no seu sentido mais genérico, não necessariamente um objeto físico mas algo que *exista*, inclusive

como uma abstração ou criação mental), um *signo* deste objeto e o possível efeito deste signo em alguma mente - o *interpretante*. Um símbolo, como na definição de Ackoff é um tipo de signo no qual esta relação triádica é estabelecida por uma convenção, lei ou costume (assim, a relação entre o termo "carro" e o objeto real carro é dada por uma convenção, um acordo, um costume dos falantes da língua portuguesa; "carro" é um símbolo do objeto carro).

Também na definição de Ackoff, ao mencionar *propriedades* somos remetidos novamente à Metafísica e à ontologia de Aristóteles (2000). Aristóteles faz uma distinção entre as categorias em Substâncias e Acidentes. Propriedades são Acidentes, dependem existencialmente das Substâncias que são as *portadoras* destes acidentes. Nestes termos, é possível supor que Ackoff, ao mencionar propriedades como *símbolos,* refere-se a valores simbólicos de propriedades, como peso, tamanho, cor, etc. Hjørland (2018) cita a definição de dados de Redman, Fox and Levitin (2017, p. 1173) da seguinte maneira:

> Within this framework, we define a datum or data item, as a triple <e, a, v>, where e is an entity in a conceptual model, a is an attribute of entity e, and v is a value from the domain of attribute a. A datum asserts that entity e has value v for attribute a. Data are the members of any collection of data items. (HJØRLAND, 2018).

Conforme esta interpretação, esses símbolos só fazem sentido ao serem referidos a uma entidade e às suas correspondentes propriedades, isto é, aos metadados de uma entidade, na forma de triplas <e, a, v>.

Floridi (2019), ao discutir a Definição Geral de Informação (GDI), vincula dados com informação: *informação = dado + semântica*. A esta questão é possível acrescentar a colocação de Ackoff: "*The difference between data and information is functional, not structural*" (ACKOFF apud ROWLEY, 2007, p. 165). Ainda segundo Floridi (2019), dado seria uma falta de uniformidade em relação a um ambiente. O autor desenvolve a GDI aprofundando três de seus aspectos: neutralidade dos relata, neutralidade taxonômica e neutralidade ontológica. A estes aspectos se poderia agregar o da existência ou não de neutralidade intencional ou de propósito. Neste sentido, constata-se que, para o tema aqui discutido, *dados são coisas artificiais, coletados/registrados com uma intencionalidade*. Não existe, em termos de *Big Data*, ou mesmo dados

científicos, dado coletado aleatoriamente ou "*data in the wild*" (FLORIDI, 2019, p. 5). "*Data in the wild*" só existe enquanto percepção (SANTAELLA, 2008, p. 96), enquanto possibilidade analítica, segundo o conceito de *Primeiridade* de C. S. Peirce (1868).

Como algo artificial, o *Big Data* que vem sendo foco de interesse na atualidade como ativo dotado de valor e, portanto, importante de ser reutilizado, é o resultado final de um processo que, segundo Floridi, começa com "*data in the wild … They are pure data or proto-epistemic data, that is, data before they are epistemically interpreted*." (FLORIDI, 2019, p. 5). Trata-se do processo que se inicia na cognição até chegar à comunicação/compartilhamento de dados (GDI.1, bem formados (GDI.2), conjunto de *símbolos* (segundo Ackoff) com um significado claramente definido (GDI.3) (Floridi, 2019). Este processo descrito por Floridi é um processo cognitivo individual, que resulta em dados prontos para serem compartilhados/comunicados socialmente e guarda analogias com a teoria da cognição de Peirce (SANTAELLA, 2008, p. 97) e a forma como os estímulos externos são percebidos e processados pelo sujeito em um *ato perceptivo*, ou suas categorias da cognição, como *Primeiridade*, *Secundidade* e *Terceiridade* (PEIRCE, 1868), até se transformarem em conhecimento pelo sujeito.

Isso posto, torna-se necessário o aprofundamento referente à semântica: Como se pode perceber a semântica de um dado? Pode-se pensar uma situação absurda/hipotética de um "*data in the wild*" *sem semântica*, como o número "62". Ontologicamente, segundo Aristóteles (2000), isto é um *acidente*, consistindo numa *quantidade*. Acidentes não existem em si, não têm independência existencial (GUARINO, 1998), ou seja, não existem sem uma instância da Categoria de *Substância* (o "e" da tripla <e, a, v>) que lhe é portadora; *quantidades* são quantidades *de* alguma coisa (o "a" da tripla <e, a, v>). Ao agregarmos semântica e esta quantidade, por exemplo, "1962 é o ano de nascimento do paciente 9698838a8fa8a01ffd5ed5c71e8e17a3", a *quantidade* "1962" (o v da tripla <e, a, v>) adquire *significado*, não uma impressão individual, mas um significado passível de ser comunicado/transferido.

No entanto, os dados que discutimos aqui não pertencem ao estágio da

cognição individual, mas pertencem a um outro, o de informação potencial sendo *comunicada/transferida* como um recurso social.

Sintetizando estas diferentes colocações de Hjørland (2018), poderíamos então dizer que: *Datum* (uma unidade de data) é a representação simbólica ou o símbolo do valor (o v da tripla <e, a, v>) de uma propriedade observável/observada (o a da tripla <e, a, v>) do fenômeno ou objeto que está sendo analisado (o e da tripla <e, a, v>). Pressupõe ou requer como pré-requisito: 1) um observador, dotado de uma intenção um fenômeno; 2) um objeto que está sendo analisado e; 3) uma proposição, declaração de uma de suas propriedades. Assim, pode-se conceber que dados são criações intencionais e artificiais.

Ao se considerar as propriedades de um fenômeno, deve-se considerar o conceito de metadados: “1962” é o ano de nascimento do paciente “9698838a8fa8a01ffd5ed5c71e8e17a3”. Nesta sentença, “1962” é o símbolo numérico da propriedade “ano de nascimento do paciente 9698838a8fa8a01ffd5ed5c71e8e17a3”. É importante definir também o que são metadados. Segundo a NISO, metadados seriam “*structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information*" (RILEY, 2017).

De maneira geral, metadados podem ser classificados em: 1) metadados de identificação (para identificar elementos, como título, resumo, autor etc.); 2) metadados estruturais (para estruturar um elemento, como a versão de um documento, índices, capítulos, páginas etc.) ou; 3) metadados gerenciais (para detalhes técnicos, como data e hora de criação, informações de acesso etc.) (PRASANNA; SASI KIRAN, 2018).

3.2 Dados não Estruturados X Dados Estruturados

Ao se considerar a relação de dados como elementos associados a semântica (FLORIDI, 2019; HØRLAND, 2018), eles só fazem sentido quando, além de integrados em triplas <e, a, v>, estão associados a outros dados e, desta forma, permitirem que sejam integrados a uma rede de conceitos e contextos.

Como visto, no nível do significado mínimo, em uma proposição formada pela tripla <e, a, v>, *v* (um dado), para ser compreendido, é inseparável de *e* e *a*. Quanto mais rica a rede de conceitos e contextos, mais semântica é agregada.

Enquanto muitos dados coletados estão estruturados mais ou menos sistematicamente segundo o esquema <e, a, v>, uma quantidade significativa é constituída de dados textuais, como documentos textuais, mensagens trocadas através das redes sociais, mensagens de correio eletrônico, etc. Tais dados são caracterizados como não estruturados e colocam dificuldades adicionais para sua gestão, processamento por máquina e reuso.

Segundo Eberendu, citando (DOAN; NAUGHTON; BAID; CHAI; CHEN; CHEN; HUANG, 2009):

> Structured data refers to data that has definite format and length, easy to store and analyze with high degree of organization. This means that the data is organized in identifiable structure to allow it response to queries to retrieve information for organizational use. (EBERENDU, 2016, p. 48)

Também citando Feldman & Sanger (2007), os mesmos autores consideram dados não estruturados como “has no particular structure. Unstructured data typically includes bitmap images/objects, text, email and other data types that are not part of a database” (EBERENDU, 2016, p. 48).

Dito de outra forma, dados estruturados podem ser entendidos como elementos ou proposições já encaixados em pré-condições, como formatos e padrões pré-definidos – ou estruturas – nos quais esses dados, então capturados ou obtidos, irão compor as narrativas a serem interpretadas. Uma vez já encaixados em estruturas, é possível obter contextos já então pré-determinados pelas próprias pré-condições ou formatos estruturantes. Por outro lado, dados não estruturados podem ser caracterizados como discursos ou elementos proposicionais não condicionadas por pré-condições ou estruturas, nos quais suas narrativas, ao serem obtidas, não se encaixam previamente nestas e encontram-se fragmentadas à espera de uma contextualização e estruturação a posteriori.

Esta questão é associada à viabilidade ou facilidade dos dados serem processados por programas. Dados estruturados como campos de uma base de dados relacional ou de uma planilha, ou dados textuais anotados, são mais fáceis

de serem processados por programas. Dados estruturados são então dados associados a metadados (uma outra questão é se estes metadados são adequados para diferentes finalidades de reuso). Para Santos e Sant'Ana (2013), uma estrutura pode ser representada através de sua relação com metadados, e representa um meio para se viabilizar a utilização e reutilização de dados. Dados estruturados são então dados que seguem um modelo ou esquema de dados prévio, como por exemplo uma tabela de uma base de dados relacional.

Para além das questões técnicas e tecnológicas do Big Data, um tópico fundamental relativo a dados estruturados e que assume um papel importante no contexto do Big Data é a *Responsible Data Science* (RDS), uma iniciativa baseada em quatro princípios *Fairness, Accuracy, Confidentiality e Transparency* (FACT) que devem ser observados no intuito de se contemplar também valores éticos socialmente aceitos e constituintes de um modelo sustentável aplicável à ciência dos dados (DATA SCIENCE CENTER EINDHOVEN, 2020) (ANDRADE *et al*., 2020).

Cumpre ressaltar que as técnicas para a análise de dados comumente são aplicadas em conjunto com as tecnologias associadas à Internet das Coisas, segundo um viés tecnicista que busca obter valor para quem as utiliza. Assim, os princípios FACT devem ser observados em conjunto aos princípios FAIR[6] (Findability, Accessibility, Interoperability e Reusability), em geral associados ao contexto de utilização de Big Data e como uma maneira de se considerar temas sociais em processos tecnológicos (ANDRADE *et al*., 2020).

Para uma melhor contextualização, a Figura 1 exibe dados de testes para Covid-19 extraídos do repositório da Fapesp/Covid19[7], cujos insumos são provenientes da USP. De acordo com a Figura, os dados são estruturados em campos de uma tabela que permitem distinguir os próprios dados, bem como seus metadados, os campos da tabela. Neste sentido, os metadados, ainda que em diferentes níveis, permitem a contextualização da narrativa contida nos dados, ocultas ou explícitas.

[6] Disponível em: https://www.go-fair.org/.

[7] Disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/100.

Figura 1 – Pacientes que fizeram teste para COVID-19

	A	B	C	D	E	F	G
1	ID_PACIENTE	IC_SEXO	AA_NASCIMENTO	CD_PAIS	CD_UF	CD_MUNICIPIO	CD_CEPREDUZIDO
2	9698838a8fa8a01ffd5ed5c71e8e17a3	F	1962	BR	SP	SAO PAULO	CCCC
3	d9fec23b3820f93a961841d569db8cb5	F	1974	BR	SP	MMMM	CCCC
4	ee507ba3a9959fdf31bca52852fd5715	F	1962	BR	SP	MMMM	CCCC
5	51590e8c53f4e8e332c05d7e6cee35c7	F	1960	BR	SP	SAO PAULO	CCCC
6	13699f0f7714fdaba277c5e360c6869c	M	1967	BR	SP	MMMM	CCCC
7	9bd839fe149857321685b1e1d8a55cbd	F	1948	BR	SP	SAO PAULO	CCCC
8	b723165a04c8e28fe2b4dcff8dc8dab3	F	1963	BR	SP	MMMM	CCCC
9	f3e55befc8f93c47d53d02a66e00bd85	M	1957	BR	SP	MMMM	CCCC
10	c4c63fb16c76294e82b33632ce4636bd	F	1960	BR	SP	SAO PAULO	CCCC

Fonte: **Repositório Fapesp/Covid 19.**

Por sua vez, dados não estruturados são dados dissociados que necessitam de metadados. Segundo vários estudos (CHAKRABORTY; PAGOLU, 2014; INMON e LINSTEDT, 2015), estes constituem o grande volume de dados tanto na web quanto nas organizações, sendo por isso necessário estruturá-los. Para além dos textos, dados não estruturados são também dados audiovisuais, como imagens estáticas, imagens em movimentos ou sons gravados. Sob quaisquer aspectos, tantos textos quanto documentos audiovisuais necessitam de representação mediante metadados, por sua própria natureza incapaz de se auto-representar. Reusar estes dados não estruturados é muito mais difícil porque, embora eles possam ser compreendidos por pessoas, são difíceis de serem compreendidos e processados por programas.

Uma distinção importante entre estas duas tipologias de dados reside na sua forma de apreensão cognitiva. Dados estruturados são apreendidos cognitivamente como um conjunto ou como o retrato ou descrição de uma situação (muitas vezes chamado de *state of affairs*). Tal descrição é formada por proposições compostas por triplas <e, a, v> como no exemplo dos dados de pacientes representados pela figura 1, ou seja, pela identificação da entidade - *e* - da qual se fazem afirmações, e de um conjunto de atributos - *a* - e valores destes atributos e dados - *v* - para esta entidade.

Para Santaella, “*Semiótica é a ciência que tem por objeto de investigação todas as linguagens possíveis, ou seja, que tem por objetivo o exame dos modos de constituição de todo e qualquer fenômeno como fenômeno de produção de significado e sentido*” (SANTAELLA, 1983, p. 13). Ela explica e integra fenômenos nas “*categorias mais universais de todas as experiências*

possíveis" (SANTAELLA, 2008, p. 7) e que vão desde as percepções individuais do mundo real que nos afetam – a *Primeiridade*, passando pela sua elaboração mental, a *Secundidade* – até se tornarem compreensíveis/inteligíveis e reconhecíveis com o aparato cultural o qual cada um de nós é dotado ao longo de sua vida cultural, ou *Terceiridade*. Isso permite sua comunicação enquanto conhecimento a outros seres humanos: "*the necessary conditions of the transmissions of meaning by signs from mind to mind*" (PEIRCE, 1931, CP 1.444).

Segundo Maciel (1974), a Lógica considera que existe uma correspondência entre pensamento e linguagem, onde as operações do pensamento corresponderiam a tipos crescentemente complexos de enunciados de linguísticos: à concepção ou *conceito* (pensamento) corresponderia o termo (linguagem), ao juízo (pensamento) corresponderia a proposição (linguagem) e ao raciocínio (pensamento) corresponderia o argumento ou inferência (linguagem). A compreensão proposta neste trabalho, a de uma proposição linguística tríplice <e, a, v>, guarda uma relação estreita com a semântica da linguagem RDF: "*RDF is an assertional language intended to be used to express propositions using precise formal vocabularies, particularly those specified using RDFS [RDF-VOCABULARY], for access and use over the World Wide Web...*" (RDF Semantics, 2004).

Em relação aos dados não estruturados, esta redução a formatos pré-concebidos não é efetivada e, assim, seus aspectos narrativos podem ser entendidos como um processo cognitivo de decodificação de mensagens, de modo a obter-se uma compreensão da narrativa ou discurso. A narrativa, implícita e oculta, encontra-se dispersa no corpo dos dados, necessitando então de uma estruturação reducionista para uma melhor extração de contextos. O contexto só é totalmente apreendido ao fim do processo, condicionado pelo caráter *sequencial* das mensagens linguísticas ou imagéticas (um vídeo) no qual o discurso/narrativa de uma situação é emitido/apreendido, conforme o exemplo de narrativa abaixo, extraído da Figura 1.

"O paciente de ID_PACIENTE = 9698838a8fa8a01ffd5ed5c71e8e17a3" é do sexo "feminino", tem ano de nascimento = "1962", nasceu no país = "Brasil",

no estado “São Paulo” e o exame foi coletado na localidade de CEP = “CCCC”. Pode-se afirmar que os dados da tabela expressa pela figura 1 formam uma descrição de uma situação ou estado de coisas.

O caráter sequencial ou procedural da linguagem natural ou imagética não permite distinção entre conteúdos e que tipo de coisa um conteúdo é, entre dados e metadados. Em mensagens digitais linguísticas ou audiovisuais, dados e metadados têm que ser identificados ou atribuídos *a posteriori*, em processos de *parsing* (análise linguística) ou decupagem de imagens (SANTOS *et al.*, 2018).

Para além dos dados, os metadados por muito tempo foram subvalorizados, ascendendo de importância sobretudo após o advento da web. Após os anos 1990, notadamente com o advento da era web, objetos não estruturados como textos, imagens, sons e vídeos se constituíram como elementos predominantes de informação. Por conta disso, os metadados passaram a desempenhar um papel fundamental como elementos associados aos dados, permitindo uma melhor contextualização destes.

Inicialmente, metadados eram inseridos automaticamente no momento da criação/captura dos dados, como vídeos, descrevendo aspectos de baixo nível (*ofness*) como texturas, cores, formas ou aspectos descritivos do próprio dispositivo de captura. Com o tempo, os metadados cresceram em importância de modo a descrever/representar contextos inseridos ou implícitos em dados não estruturados. Além de inseridos automaticamente, eles também passaram a ser atribuídos, descrevendo ou representando aspectos de alto nível (*aboutness*) como os discursos ou narrativas contidos nos dados. De acordo com Fisher e Sheth (2004), os metadados podem ser agrupados segundo seus tipos. Os principais podem ser concebidos como:

- Metadados sintáticos: proveem informações genéricas e automáticas, geralmente adicionados ou inscritos no momento da criação do objeto;

- Metadados semânticos: são metadados associados, implícita ou explicitamente a um conteúdo, cuja relevância é determinada pela sua posição ontológica dentro de um determinado domínio ou aplicação.

Os metadados aumentam sua importância e tipologia de acordo com o

grau de estruturação dos dados. Também permitem descrições/representações de maneira alfabética ou numérica. Assim, quanto mais os dados forem de natureza não estruturada, maior a necessidade de metadados estruturais e semânticos para contextualizá-los e melhor representar conteúdos e narrativas ocultas ou não explícitas nestes. Assim, quanto mais os dados forem de natureza não estruturada, maior a necessidade de metadados estruturais e semânticos para contextualizá-los e melhor representar conteúdos e narrativas ocultas ou não explícitas nestes.

4 VOCABULÁRIOS E AS TECNOLOGIAS DE DADOS ABERTOS INTERLIGADOS

As tecnologias de Dados Abertos Interligados são parte integrante do projeto da *Web* Semântica, ou *Web* de dados. Este projeto foi inicialmente formulado pelo cientista da computação Tim Berners-Lee, o criador da *Web*, entre outros. A *Web* Semântica tem como objetivo propor "*A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*" (BERNERS-LEE, HENDLER, LASSILA, 2001). Ainda segundo os autores "*Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully*" (BERNERS-LEE, HENDLER, LASSILA, 2001). A Web Semântica então *"... will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users*" (BERNERS-LEE, HENDLER, LASSILA, 2001).

A *Web* Semântica se refere a conteúdos representados de forma que possam ser "compreendidos" tanto por máquinas quanto por pessoas. A *Web* atual, podendo ser considerada como uma web sintática, é formada por páginas formatadas em linguagem com base *Hypertext Markup Language* (HTML), acessíveis e interligadas umas com as outras através de *links* lidos por navegadores. Neste sentido, o HTML é uma linguagem de marcação de conteúdos, com um conjunto pré-definido de marcações que instruem os navegadores a exibi-las na tela dos computadores para os usuários humanos. O conteúdo das páginas em HTML é interpretado pelos navegadores para se tornar

legível e visualmente interpretável por pessoas.

A proposta da Web Semântica é diferente: a Web não seria mais constituída por páginas para serem lidas por pessoas, mas por recursos informacionais representando coisas concretas, como pessoas, produtos, monumentos, acidentes geográficos, por exemplo, ou coisas abstratas, como um gênero musical, uma disciplina científica, ou somente uma existência de objeto digital, como uma foto em *.jpg* ou um artigo científico em *.pdf*. Cada um destes recursos é identificado univocamente por um link, um *Uniform Resource Identifier* (URI). Um recurso, identificado/acessado por seu URI, é descrito por conjuntos de propriedades e valores destas propriedades, por exemplo:

"A página http://www.uff.br tem como autor João da Silva".

Tem-se então uma afirmação que consta de três elementos: o sujeito, "http://www.uff.br", o predicado, "tem como autor" e o objeto, "João da Silva". Este modelo de descrição de recursos é formado por afirmações linguísticas constituídas por triplas <Sujeito> <Predicado> <Objeto> é o RDF - *Resource Description Framework* (RDF Primer, 2004).

Este modelo pressupõe uma semântica mínima, derivada de sua correspondente afirmação linguística, ou seja, são claramente identificados e aparecem nesta ordem o Sujeito, o Predicado e o Objeto que formam a tripla. Conjuntos de triplas com o mesmo Sujeito descrevem um mesmo recurso. A semântica mínima do modelo RDF permite processar conjuntos de triplas onde um ou dois dos Sujeito(s), Predicado(s) ou Objeto(s) são desconhecidos, como:

- Quem é o autor da página http://www.uff.br? < http://www.uff.br > < tem como autor> < ???>.

- Que papel João da Silva tem em relação à página http://www.uff.br? < http://www.uff.br > < ??? > < João da Silva >.

- Quais são todas as afirmações sobre a página http://www.uff.br? < http://www.uff.br > < ??? > < ??? >.

O SPARQL - *SPARQL Protocol and RDF Query Language* - é a linguagem de consulta que permite consultar conjuntos de triplas RDF, materializando assim a proposta da Web Semântica de consultar a Web como se fosse uma base de dados.

O RDF pode ser representado (serializado em linguagem técnica de computação) em diversos formatos, como RDF/XML, N-Triples, JSON, TURTLE (RDF Primer, 2004). Ainda que triplas RDF representadas nestes formatos não sejam amigáveis nem claramente legíveis para pessoas como as páginas HTML, elas contêm elementos que permitem a navegadores que compreendam estes formatos a exibi-las de maneira amigável para pessoas, quando for o caso. Isso porque o objetivo principal dos recursos descritos em RDF é permitir o processamento por máquinas, ajudando assim a organizar, recuperar, tornar acessíveis estes recursos.

Naturalmente que dada uma tripla como < http://www.uff.br > < tem como autor> < João da Silva >, uma máquina não pode fazer muito mais que identificar quem é o Sujeito, o Predicado ou o Objeto da tripla; neste exemplo Predicado e Objeto são nomes, sequências de caracteres só compreensíveis por pessoas, detentoras de um conjunto de informações contextuais e culturais, acumuladas ao longo de suas histórias de vida.

Neste contexto, então, entra o papel dos vocabulários. Enquanto os recursos previstos na web Semântica, representados em linguagens de marcação como XML, RDF, HTML, etc. são *conteúdos*, programas são *procedimentos*. Programas só processam conteúdos e, para isso, precisam ser instruídos (programados) claramente sobre o que fazer com determinado conteúdo em determinada situação. Por seu turno, os Vocabulários a serem usados na Web Semântica, ou LOV - *Linked Open Vocabularies* - (ZENG, 2019b), definem, restringem e especificam claramente a Semântica dos seus conceitos. Por exemplo, o vocabulário de metadados DC (*Dublin Core*) - define claramente a semântica de um dos seus conceitos ou elementos, como o *dc:creator*, que consiste no criador/autor ou responsável por um recurso, por exemplo, um artigo científico digital; mais do que isto, o elemento *dc:creator* possui, ele próprio, um identificador persistente unívoco, um *link*, um URI, http://purl.org/dc/elements/1.1/creator. Este identificador persistente unívoco permite que um programa, para processar este elemento do vocabulário de metadados DC, seja desenvolvido de forma clara e inequívoca, a partir da semântica especificada e padronizada no vocabulário DC.

Num outro exemplo, sejam as seguintes triplas RDF:

- <libro0237> <title> <Dom Quixote> <http://catalogo.bne.es/libro0237> <http://purl.org/dc/elements/1.1/title> <Dom Quixote>; e
- <emp0027> <title> <Presidente> <http://www.company.com/0027> <http://www.w3c.org/2006/vcard/ns/title> <Presidente>.

Os predicados de ambas as triplas aparentemente são idênticos, *title*, só diferem pelo *link* para o vocabulário; no primeiro exemplo é http://purl.org/dc/elements/1.1/title e no segundo é http://www.w3c.org/2006/vcard/ns/title. Estes *links* para vocabulários diferentes, também chamados de *namespaces* - espaços de nomes - permitem aos programas que processam as triplas identificar univocamente os diferentes conceitos nos diferentes vocabulários que servem de predicados para as duas triplas e mesmo processar as duas triplas simultaneamente, sem confundir suas semânticas, porque elas não se restringem ao eventual significado informal de *title*, mas sim, a este significado *no contexto* (*namespaces*) dos vocabulários DC ou Vcard.

Numa variante do primeiro exemplo, seja a seguinte tripla:

- <Dom Casmurro> <Autor> < Machado de Assis>.

Neste caso, ao invés do Sujeito, Predicado e Objeto da tripla em linguagem natural, sujeitos a ambiguidades, interpretações e erros de digitação, pode-se utilizar o URI para identificar cada um deles:

<http://acervo.bndigital.bn.br/sophia/index.asp?codigo_sophia=4883>
<http://purl.org/dc/elements/1.1/creator>
<https://viaf.org/viaf/95151633/>.

O primeiro URI é um *link* para o registro da primeira edição da obra *Dom Casmurro* e de sua cópia digital, no acervo da *BN Digital*; o segundo é um *link* para o elemento *dc:creator* no vocabulário de metadados DC, já mencionado anteriormente; e o terceiro é um *link* para o registro de *Machado de Assis* no vocabulário ou base de dados VIAF - *Virtual International Authority File*, uma iniciativa internacional para padronização e identificação unívoca de autoridades, pessoas e instituições.

É desta maneira, através de identificadores únicos e persistentes, que

vocabulários de metadados e de dados podem ser usados para atribuir semântica a predicados e objetos de triplas em RDF. Muitos antigos Vocabulários, como os Tesauro da UNESCO[8], o Tesauro da *FAO*[9], o *AGROVOC Thesaurus*[10], os Vocabulários da *Fundação Paul Getty*[11], o *Art and Architecture Thesaurus*, o *Union List of Artists Names*, o *Cultural Objects Name Authority*, o *Getty Thesaurus of Geographic Names*, os DeCS/MeSH, Descritores em Ciência da Saúde[12], o LCSH - *Library of Congress Suject Headings*[13], além de muitos outros, estão sendo reestruturados para serem compatíveis com as tecnologias de Dados Abertos Interligados (SOERGEL *et al*, 2004; MACULAN, 2015).

Nas recomendações internacionais, para que dados sejam considerados abertos, há uma qualificação de 1 a 5 estrelas, onde a quarta e quinta estrela é atribuída quando dados são disponíveis em formato RDF, com predicados e objetos referidos a vocabulários padronizados e largamente reconhecidos pela comunidade em determinado domínio e interligados uns com os outros para fornecer um rico contexto. Com relação aos dados de pesquisa, tema que vem sendo objeto de atenção crescente e de políticas públicas em níveis nacionais e internacionais, a iniciativa *GO FAIR* recomenda um conjunto de princípios para a publicação de dados de pesquisa de modo a que eles tenham os atributos de *Findability, Accessibility, Interoperability, Reuse*. Para que dados de pesquisa alcancem estes atributos, os dados devem poder ser acessados através de um URI e sua representação deve ser formatada em RDF, usando os LOV vocabulários.

A ideia por trás dos princípios FAIR é permitir que dados de pesquisa possam ser tratados por máquinas, o princípio M4M – *Metadata for Machines*. Um exemplo concreto e dramático da importância dos dados de pesquisa e da adoção de princípios que permitam sua ampla disseminação e reuso é o formulário de coleta de dados sobre a epidemia de COVID-19 proposto pela

[8] Disponível em: http://vocabularies.unesco.org/browser/thesaurus/en/
[9] Disponível em: http://aims.fao.org/aos/agrovoc/c_8003.html
[10] Disponível em: https://agrovoc.fao.org/browse/agrovoc/en/
[11] Disponíveis em: https://www.getty.edu/research/tools/vocabularies/lod/
[12] Disponível em: https://decs.bvsalud.org/ths/
[13] Disponível em: https://id.loc.gov/authorities/subjects.html

OMS - *Organização Mundial da Saúde*. A iniciativa *GO FAIR* propõe a criação de uma rede mundial de catálogos que façam referência a dados de pesquisa depositados em repositórios e que estejam disponíveis segundo os princípios FAIR, os *FAIR Data Points*; o Brasil, inclusive, participa desta iniciativa através do a inciativa VODAN-Br - *Virus Outbreak Data Network*. Todos os campos do formulário devem ser preenchidos com metadados e dados referenciados a vocabulários, de modo a permitir sua padronização sem a qual não seria possível seu processamento por máquinas.

Vocabulários, para serem usados como Dados Abertos Interligados, precisam atender a requisitos como, por exemplo, ter seus conceitos identificados de forma persistente e unívoca através de URI válidos na internet como um todo, serem representados em formatos legíveis por máquinas como RDF, conter definições precisas de sua semântica e, geralmente, são multilíngues. Muitos destes Vocabulários que atendem aos princípios dos Dados Abertos Interligados podem ser encontrados no serviço de registro de vocabulários LOV já mencionado.

É desta forma, atendendo aos requisitos de poderem ser usados com os Dados Abertos Interligados como descrito acima, que Vocabulários, uma área de estudos, pesquisa e utilização prática da CI/OC pode contribuir para endereçar a questão do *Big Data*.

5 CONSIDERAÇÕES FINAIS

É cabível à CI/OC exercer um papel de destaque no entorno do fenômeno *Big Data*, favorecendo o potencial de reuso dos dados; este papel se dá principalmente através da manipulação dos vocabulários (de metadados e de dados), cujos princípios teóricos, metodológicos e práticos estas áreas já desenvolvem há tempos.

Dados" (plural, "data") é distinguido de dado (singular, "datum", o que seria uma unidade de dados), assim como os conceitos de metadado X dado - representações de propriedades de objetos, eventos e seu ambiente.

Dados enquanto símbolos ou representações só fazem sentido ao serem referidos à uma entidade e às suas correspondentes propriedades, ou seja, à

metadados de uma entidade, triplas <e, a, v>.

Sintetizando poderíamos então dizer que: Datum (uma unidade de data) é (a representação simbólica) o símbolo do valor (o v da tripla <e, a, v>) **de** *uma* propriedade observável/observada (o a da tripla <e, a, v>) **do fenômeno ou objeto que está sendo analisado** (o e da tripla <e, a, v>). Pressupõe um observador, com uma intenção um fenômeno ou objeto que está sendo analisado e uma proposição declarando de uma de suas propriedades. Dados são criações intencionais e artificiais.

Dados em uma definição informal se referem a conjuntos de proposições descrevendo as propriedades e valores (estes sim, os dados) destas propriedades para uma entidade.

Dados estruturados se referem a descrições de um estado de coisas nas quais uma entidade é descrita através de um conjunto de suas propriedades e valores, que podem ser observadas em sua totalidade, como no exemplo da Figura 1.

Dados não estruturados são também descrições de um estado de coisas de carácter processual, onde só no decorrer do processo um observador pode perceber a entidade descrita e suas propriedades e valores como um todo; exemplos típicos são um texto em linguagem natural, um vídeo.

Este trabalho, em suma, visa oferecer uma contribuição ao se aproximar do fenômeno *Big Data* a partir de um fenômeno mais conhecido pela CI/OC, que são os dados. Insere, assim, dados em um esquema conceitual inseparável da entidade que está sendo descrita/representada, de suas propriedades descritivas (metadados) e dos valores destas propriedades (estes sim, os dados). A partir desta conceituação, distingue dados estruturados de dados não estruturados. Ressalta ainda o papel dos vocabulários de metadados e dados para atribuir semântica acordada e padronizada às propriedades e seus valores de uma entidade representada em ambiente digital para que máquinas possam processar e fazer inferências sobre estas representações.

Como limitações do artigo e programa para futuras pesquisas, entende-se como necessária uma maior exploração das diferenças entre os vocabulários a serem usados por máquinas dentro do entorno da Web Semântica e os SOC

tradicionais desenvolvidos pela CI/OC como instrumentos semânticos auxiliares para a recuperação de informações. Vocabulários não são a única maneira de atribuir semântica computacional a dados digitais; dados não estruturados textuais, em especial, vêm sendo processados por técnicas de Processamento de Linguagem Natural (PLN) e de reconhecimento de entidades (em inglês NER - *Named Entity Recognition*[14]), dentre outras tecnologias de inteligência artificial. Assim, torna-se necessário também considerar questões sobre como exatamente um vocabulário pode atribuir semântica para máquinas.

REFERÊNCIAS

ANDRADE, M. C.; GONÇALEZ, P. R. V. A.; BERTI JUNIOR, D. W; BAPTISTA, A. A.; CONEGLIAN, C. S. Responsible data science: Impartiality, accuracy, confidentiality and transparency of data. **Informação & Informação,** Londrina, v. 25, n. 2, p. 26-48, 2020.

ARISTÓTELES. **Categorias.** Lisboa: Instituto Piaget, 2000.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American,** may, 2001.

BUNGE, M. **Treatise on Basic Philosophy:** Volume 3 Ontology I: The furniture of the World. Dordrecht, Holland, Boston, USA: D Reidel Publishing, 2015.

CAPURRO, R. Angeletics - A Message Theory. *In*: DIEBNER, H. H.; RAMSAY, D. L. (ed.). **Hierarchies of Communication.** Karlsruhe: ZKM - Center for Art and Media, 2003.

CHAKRABORTY, G.; PAGOLU, M. Analysis of Unstructured Data: Applications of Text Analytics and Sentiment. *In*: SAS GLOBAL FORUM, 8., 2014, Washington DC. **Conference Paper** […]. SAS: Washington DC, Mar. 2014.

CLARKE, S. G. D. The Information Retrieval Thesaurus. **KO KNOWLEDGE ORGANIZATION,** [*S. l.*], v. 46, n. 6, p. 439-459, 2019.

COGNIZANT. **Making Sense of Big Data in the Petabyte Age.** Cognizant, *20-20 insights*, jun. 2011. Disponível em: https://www.cognizant.com/whitepapers/Making-Sense-of-Big-Data-in-the-Petabyte-Age.pdf. Acesso em: 02 abr. 2021.

[14] Ver em https://en.wikipedia.org/wiki/Named-entity_recognition

DATA SCIENCE CENTER EINDHOVEN. **Responsible Data Science:** Ensuring fairness, accuracy, confidentiality & transparency by design. 2020. Disponível em: https://www.tue.nl/en/research/research-areas/data-science/responsible-data-science/. Acesso em: 02 dez. 2020.

EBERENDU, A. C. Unstructured Data: an overview of the data of Big Data. **International Journal of Computer Trends and Technology,** v. 38, n. 1, p. 46-50, 2016.

FAIR Compliant Biomedical Metadata Templates. CEDAR, Center for Expanded Annotation and Retrieval, University of Stanford, Department of Medicine, 2019. Disponível em: https://medicine.stanford.edu/2019-report/cedar-to-the-rescue.html. Acesso em: 15 ago. 2021.

FISHER, M.; SHETH, A. Semantic Enterprise Content Management. *In*: SINGH, M. P. **The practical handbook of internet computing.** (Computer and Information Science Series). Boca Raton, FL: Chapmann & Hall/CRC, 2004.

FLORIDI, L. Semantic Conceptions of Information. *In*: ZALTA, E. N. (ed.). **The Stanford Encyclopedia of Philosophy.** Palo Alto: Metaphysics Research Lab, 2019. Disponível em: https://plato.stanford.edu/archives/win2019/entries/information-semantic/. Acesso em: 21 dez. 2019.

GUARINO, N. Some ontological principles for designing upper level lexical resources. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 1., 1998. Granada. **Proceedings** […]. Granada: ELRA, 1998. Disponível em: https://arxiv.org/pdf/cmp-lg/9809002. Acesso em: 22 maio 2005.

HEY, T.; TREFETHEN, A. *The data deluge*: An e-science perspective. *In*: BERMAN, F.; FOX, G. C.; HEY, A. J. G. (ed.). **Grid computing:** making the global infrastructure a reality. Wiley: West Sussex, 2003. p. 809-824. Disponível em: https://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf. Acesso em: 10 out. 2020.

HJØRLAND, B. Data (with big data and database semantics). **Knowledge Organization,** v. 45, n. 8, p. 685-708, 2018.

INMON, W.; LINSTEDT, D. **Data Architecture**: a primer for the data scientist. Waltham, MA, Elsevier, 2015.

MACIEL, J. **Elementos de Teoria Geral dos Sistemas.** Petrópolis: Vozes, 1974.

MARCONDES, C. H. Em Busca de uma Semântica do Digital, Ou “As They May Think”. **PontodeAcesso,** Salvador, v. 6, n. 2, p. 35-73, 2012.

ORILIA, F.; PAOLETTI, M. P. Properties. *In*: ZALTA, E. N. (ed.). **The Stanford Encyclopedia of Philosophy.** Palo Alto: Metaphysics Research Lab, 2020.

Disponível em: https://plato.stanford.edu/archives/win2020/entries/properties/. Acesso em: 9 maio 2020.

PEIRCE, C. S. **Collected papers of Charles Sanders Peirce:** principles of philosophy. Cambridge: Harvard University Press, 1931. v. 1.

PEIRCE, C. S. On a new list of categories. *In:* AMERICAN ACADEMY OF ARTS AND SCIENCES*,* 7., 1868, Cambridge. **Proceedings** […]. American Academy of Arts and Sciences: Cambridge, 1868. p. 287-298. Disponível em: http://www.bocc.ubi.pt/pag/peirce--charles-list-categories.pdf. Acesso em 28 jul. 2018.

PRASANNA, J. K. L.; SASI KIRAN, K. S. M. Significance of metadata and data modelling of metadata by using mark logic. **International Journal of Engineering and Advanced Technology,** v. 8, n. 2, p. 76-78, 2018.

RAYWARD, W. B. **The Universe of Information:** the work of Paul Otlet for Documentation and international organization. Moscou: FID/VINITI, 1975.

RDF Semantics. W3C, 2004. Disponível em: https://www.w3.org/TR/rdf-mt/. Acesso em: 27 ago. 2021.

RILEY, J. **Understanding metadata**: what is metadata and what is it for: a primer. Baltimore: NISO, 2017. Disponível em: https://groups.niso.org/apps/group_public/download.php/17446/Understandin Metadata.pdf. Acesso em: 13 mar. 2021.

ROWLEY, J. The wisdom hierarchy: representations of the DIKW hierarchy. **Journal of information science,** v. 33, n. 2, p. 163-180, 2007. Disponível em: http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/rowleydikw.pdf. Acesso em: 14 jun. 2013.

SANTAELLA, L. Epistemologia semiótica. **Cognitio: Revista de Filosofia,** v. 9, n. 1, p. 93-110, 2008. Disponível em: https://revistas.pucsp.br/cognitiofilosofia/article/viewFile/13531/10042. Acesso em: 12 nov. 2020.

SANTAELLA, L. **O que é Semiótica.** São Paulo: Ed. Brasiliense, 1983.

SANTOS, F. E. P.; FARIAS, M. G. G.; FEITOSA, L. T.; CAVATI SOBRINHO, H. Definição de metadados e critérios de indexação para documentário em repositório audiovisual. **Revista Brasileira de Biblioteconomia e Documentação,** v. 14, n. 3, p. 237-261, 2018. Disponível em: https://rbbd.febab.org.br/rbbd/article/viewFile/1092/1089. Acesso em: 19 nov. 2020.

SANTOS, P. L. V. A. C. SANT'ANA, R. C. G. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação. **Ciência da Informação,** Brasília, v. 42, n. 2, p. 199-209, maio/ago. 2013.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. **Informação & Informação,** Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016.

MACULAN, B. C. M. S. **Estudo e aplicação de metodologia para reengenharia de tesauro:** remodelagem do THESAGRO. 2015. 345 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas gerais, Belo Horizonte, 2015. Disponível em: https://repositorio.ufmg.br/bitstream/1843/BUBD-9ZKMUV/1/maculan_tese_arq_final.pdf.pdf. Acesso em: 24 maio 2019.

SOERGEL, D.; LAUSER, B.; LIANG, A.; FISSEHA, F.; KEIZER, J.; KATZ, S. Reengineering thesauri for new applications: the AGROVOC example. **Journal of digital information,** v. 4, p. 1-23, 2004. Disponível em: http://hdl.handle.net/10760/15694. Acesso em: 25 abr. 2016.

ZENG, M. L. Interoperability. *In:* HJØRLAND, B.; GNOLI, C. (ed.). ISKO **Encyclopedia of Knowledge Organization.** ISKO, 2019a. Disponível em: http://www.isko.org/cyclo/interoperability. Acesso em: 18 set. 2019.

ZENG, M. L. Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article. **El profesional de la información,** v. 28, n. 1, 2019b. Disponível em: https://doi.org/10.3145/epi.2019.ene.03. Acesso em: 22 jan. 2019.

THE ROLE OF VOCABULARIES TO THE ACCESS AND REUSE OF BIG DATA

ABSTRACT

Objective: Similar to the "information explosion", the Big Data phenomenon has been increasingly the object of CI/OC. How to discover, access, process and reuse the huge and growing amount of data that is continuously made available on the web by our society? In particular, how to deal with the so-called "unstructured data", textual documents, which have always been the object of CI/OC? **Methodology**: Broad spectrum theories such as Ontology and Semiotics were used to analyze data as an essential element of Big Data, especially "unstructured data". **Results**: From the analysis of several data definitions, a given is identified as part of already known logical and semiotic schemes, the propositions. One piece of data is found together with others, forming data sets. Data sets are actually sets of propositions. These are present in what is known as structured data - tables in relational databases or spreadsheets. Textual documents also contain sets of propositions. Structured data is compared to "unstructured data". **Conclusions**: Although at the limit, both contain propositions and can be equivalent, as sets, structured data are expressed and perceived as a whole, sets of "unstructured data" are procedural, expressed sequentially, which makes the identification of unstructured data more difficult in text documents for processing by machines.

Descriptors: Big Data. Vocabularies. Structured data. Unstructured data. Linked Open Data.

EL ROL DE LOS VOCABULARIOS PARA EL ACCESO Y LA REUTILIZACIÓN DE LOS BIG DATA

RESUMEN
Objetivo: Similar a la “explosión de la información”, el fenómeno de Big Data ha sido cada vez más objeto de CI / OC. ¿Cómo descubrir, acceder, procesar y reutilizar la enorme y creciente cantidad de datos que nuestra sociedad pone continuamente a disposición en la web? En particular, ¿cómo tratar los denominados “datos no estructurados”, documentos textuales, que siempre han sido objeto de CI / OC? Metodología: Se utilizaron teorías de amplio espectro como la ontología y la semiótica para analizar los datos como elemento esencial del Big Data, especialmente los “datos no estructurados”. **Resultados**: A partir del análisis de varias definiciones de datos, se identifica un dato como parte de esquemas lógicos y semióticos ya conocidos, las proposiciones. Un dato se encuentra junto con otros, formando conjuntos de datos. Los conjuntos de datos son en realidad conjuntos de proposiciones. Estos están presentes en lo que se conoce como datos estructurados: tablas en bases de datos relacionales u hojas de cálculo. Los documentos textuales también contienen conjuntos de proposiciones. Los datos estructurados se comparan con los "datos no estructurados". **Conclusiones**: Aunque en el límite, ambos contienen proposiciones y pueden ser equivalentes, como conjuntos, los datos estructurados se expresan y perciben como un todo, los conjuntos de "datos no estructurados" son procedimentales, expresados secuencialmente, lo que dificulta la identificación de datos no estructurados en documentos de texto para procesamiento por máquinas.

Descriptores: Big Data. Vocabularios. Datos estructurados. Datos no estructurados. Datos abiertos enlaçados.

ANEXO 1 – Resultados da Pesquisa Bibliográfica

Estratégia de busca: Google Scholar: "big data"[title] "information science" "knowledge organization" definition "big data is" from 2010 to 2021, no citations, no patents.

Autor(es)	Título	Ano	Fonte
A David, N Ndjock	Big data, Knowledge Organization and decision making: opportunities and limits	2018	https://books.google.com/books?hl=en&lr=&id=MyDDwAAQBAJ&oi=fnd&pg=PA95&dq=%22big+data%22+%22information+science%22+%22knowledge+organization%22+definition+%22big+data+is%22&ots=MUMbxfhJJM&sig=yA3T6Cc3DMtee56HNtxA26ZKyk
P Ajibade, SM Mutula	Big Data Research Outputs in the Library and Information Science: South	2020	http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=07954778&AN=143392890&h=WQbXfqCOAzpZnhxkoawm

	African's Contribution using Bibliometric Study of Knowledge Production.		Xp2wOCt2xd7d8mecFXC%2B9YD%2BDGDA6h4Bce0npl7NJx7xvWwrAr9%2F9E9v4makffjP9g%3D%3D&crl=c
B Hjørland	Data (with big data and database semantics)	2019	https://www.nomos-elibrary.de/10.5771/0943-7444-2018-8-685/data-with-big-data-and-database-semantics-volume-45-2018-issue-8
A Shiri	Linked Data Meets Big Data: A Knowledge Oraganization Systems Perspective. The ASIS&T Special Interest Group on Classification Research (SIG/CR) …	2014	https://era.library.ualberta.ca/items/2edaa5e0-3034-448e-bb5a-1b931f84ff78/view/643d8d9b-6280-4e28-aa0e-6fedc2a23a9e/ACRO_24_1_16.pdf
TS Calvard	Big data, organizational learning, and sensemaking: Theorizing interpretive challenges under conditions of dynamic complexity	2016	https://journals.sagepub.com/doi/abs/10.1177/1350507615592113
YS Hu	The impact of increasing returns on knowledge and big data: from Adam Smith and Allyn Young to the age of machine learning and digital platforms	2020	https://www.jstor.org/stable/10.13169/prometheus.36.1.0010
A Shiri	Making sense of big data: A facet analysis approach	2014	https://www.nomos-elibrary.de/10.5771/0943-7444-2014-5-357/making-sense-of-big-data-a-facet-analysis-approach-volume-41-2014-issue-5

Recebido em: 30.09.2021
Aceito em: 29.11.2021