XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

GT 2 – Organização e Representação do Conhecimento

EXPRESSIVIDADE SEMÂNTICA DO *BIG DATA*: EXTRAINDO INFORMAÇÕES DE DADOS

SEMANTIC EXPRESSIVITY OF BIG DATA: EXTRACTING INFORMATION FROM DATA

Durval Vieira Pereira – Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Linair Maria Campos – Universidade Federal Fluminense (UFF)
Sergio de Castro Martins – Universidade Federal do Rio de Janeiro
Mauricio Augusto Cabral Ramos Jr.
Carlos Henrique Marcondes – Universidade Federal Fluminense (UFF) e Universidade Federal de Minas Gerais

Modalidade: Trabalho Completo

Resumo: Estuda a relação entre dados digitais e sua expressividade semântica, no contexto do *Big Data*. Problematiza os dados enquanto recurso semântico, que atingem todo seu potencial quando processados pelas tecnologias da informação. Objetiva examinar como o conceito de expressividade semântica atua no âmbito do *Big Data*, no que tange a extração de informações de dados. Conceitua dados enquanto artefatos criados a partir de uma intencionalidade, entes semióticos, representações de fenômenos ou entidades do mundo real. Sugere que dados se organizam em diferentes níveis, desde o mais simples dado digital até os agregados mais complexos, formando conjuntos ou sistemas de dados, tornando os dados mais expressivos e capazes de gerar semântica para humanos e para máquinas. Classifica-se como pesquisa de natureza qualitativa, de tipo exploratório, com caráter teórico-conceitual e procedimento bibliográfico. Apresenta como resultado um quadro onde os diferentes tipos de agregados de dados são organizados em níveis crescentes de complexidade. À medida que se tornam mais complexos estes agregados tornam-se mais expressivos semanticamente e, ao serem processados, podem gerar semântica/informação para humanos e para máquinas. Conclui que, em um ambiente de *Big Data,* a "semântica" emerge dos dados segundo dois eixos: o eixo das representações ou modelagem conceitual (Organização do Conhecimento), e eixo do processamento destas representações, a modelagem estatística (Ciência de Dados). Os resultados mostram que pode haver uma aproximação entre a Organização do Conhecimento e *Big Data* e Ciência de Dados.

Palavras-chave: Big Data; Expressividade semântica; Teoria dos Níveis Integrativos; Organização do Conhecimento; Ciência de Dados

Abstract: It studies the relationship between digital data and its semantic expressiveness, in the context of Big Data, seeking to bring Knowledge Organization and Data Science closer together. It problematizes data as a semantic resource, which only reaches its full potential when processed by information technologies. It aims to examine how the concept f semantic expressiveness operates within the scope of Big Data, in terms of removing information from data. It conceptualizes data as artifacts created from intentionality, semiotic entities, representations of phenomena or entities in the real world. It suggests that data is organized at different levels, from the simplest digital data to the most complex aggregates, forming data sets or systems, making data more expressive and capable of generating semantics for humans and machines. It is classified as qualitative research, exploratory in nature, with a theoretical-conceptual character and bibliographic procedure. The result is a table where the different types of data aggregates are organized at increasing levels of complexity. As they

become more complex, these aggregates become more semantically expressive and, when processed, can generate semantics/information for humans and machines. It concludes that, in a Big Data environment, “semantics” emerges from data along two axes: the axis of representations or conceptual modeling (Knowledge Organization), and the axis of processing these representations, statistical modeling (Data Science). The results show that there may be a rapprochement between Knowledge Organization and Big Data and Data Science.

Keywords: Big Data; Semantic expressiveness; Integrative Levels Theory; Knowledge Organization; Data Science

1 INTRODUÇÃO

Decidiu-se iniciar esta seção com uma citação de Latour (2000), pois ela demonstra como a superposição de dados representando fenômenos ou entidades diferentes pode ser altamente informativa, revelando *insights* e relações antes não percebidas entre eles.

> [...] Marey, o grande fisiologista (e inventor do inverso do cinema!), pôde sobrepor o mapa da Rússia, a medida das temperaturas, o percurso da *Grande Armée*, a data da seus deslocamentos e, mais tragicamente, o número de soldados sobreviventes em cada bivaque! (Latour, 2000, p. 30).

Na sociedade contemporânea, as atividades humanas são mediadas pelas tecnologias de informação, o chamado processo de datificação (Mejias; Couldry, 2019). Este é o processo de se transformar informações geradas por estas atividades em dados digitais, possibilitando a coleta, o armazenamento, a análise e o uso desses com finalidades específicas e a geração de novas informações. O processo gera uma quantidade extremamente grande de dados, fenômeno denominado *Big Data*, o qual vem se tornando fundamental nas organizações por apresentar um elevado potencial semântico.

Esta pesquisa trata dados digitais enquanto recursos semânticos. Estes, ao representarem entidades ou fenômenos da realidade, só atingem todo seu potencial quando processados pelas tecnologias da informação. Isto apresenta desafios como a questão dos vieses embutidos em grandes conjuntos de dados (HARE, 2023), a interoperabilidade semântica – o intercâmbio de dados entre diferentes sistemas sem perda de sua semântica original -, a construção de instrumentos de padronização semântica de dados como ontologias, vocabulários e padrões de metadados processáveis por máquinas.

A questão do *Big Data* é ampla e suscita muitas discussões como as apontadas acima. Embora dados sejam hoje um subproduto corriqueiro nas mais variadas atividades humanas, respostas mais científicas, menos simplórias, ou circulares com outros fenômenos como informação e conhecimento, precisam ser dadas para questões como: dentro do espectro das

tecnologias da Web Semântica, o que são dados? Qual o potencial semântico dos dados? Como é gerada a semântica (para humanos e para máquinas) a partir de dados digitais?

Estas questões são o foco da pesquisa. A pesquisa sugere que dados se organizam em diferentes níveis, desde o mais simples dado até agregados (Barreto, 2008; Foskett, 1961) mais complexos, formando conjuntos ou sistemas de dados, como campos, registros, tabelas, modelos conceituais, ontologias; à medida que se tornam mais complexos, em maior volume e são processados, estes agregados de dados nos diferentes níveis tornam-se mais expressivos e podem gerar semântica/informação para humanos e para máquinas. O trabalho tem o objetivo de discutir teoricamente esta hipótese, propor uma conceituação da questão e mostrar indícios de sua factibilidade.

2 PROCEDIMENTOS METODOLÓGICOS

Trata-se de uma pesquisa em andamento, de natureza qualitativa, de tipo exploratório; tem carácter teórico-conceitual e procedimento bibliográfico. Apoia-se nas bases teóricas da Organização do Conhecimento e Ciência de Dados. Utiliza ainda a Teoria dos Níveis Integrativos, a Semiótica e a Ontologia para analisar dados e seus agregados enquanto representações (portanto, criações artificiais, intencionais) dos mais diferentes fenômenos, capazes de gerar semântica, ou seja, com potencial de informar pessoas e a sociedade.

3 DADOS E *BIG DATA* NA ORGANIZAÇÃO DO CONHECIMENTO E CIÊNCIA DE DADOS

Esta discussão é desenvolvida no contexto do "*Big Data"*, passando pelos autores dentro da Organização do Conhecimento e da Ciência de Dados que discutem as relações entre dado e níveis crescentes de semântica.

Hjørland (2018) menciona que "today, there is much talk about the data deluge, and the necessity to deal with it in various fields including computer science and information science, just as there are tendencies to establish a new field, 'data science'". A área de Ciência de Dados tem se consolidado como "campo de estudo que se destaca pela capacidade de auxiliar a descoberta de informação útil a partir de grandes ou complexas bases de dados, bem como a tomada de decisão orientada por dados" (Fundação Oswaldo Cruz, 2019, p. 1).

As ações humanas expressadas na *Web*, bem como diversos dispositivos e sensores geram registros, que podemos entender como dados. A quantidade massiva de dados na internet consolidou-se como o que hoje se entende como *Big Data*, isto é, a geração de dados

em quantidade, variedade e velocidade sem precedentes na história, armazenadas em servidores na grande rede mundial, popularmente rotulada como “nuvem”.

A empresa Cognizant (2011, p. 1) em seu *newsletter* afirma: “While data volume proliferates, the knowledge it creates has not kept pace”. Ou seja, chama a atenção para uma questão que vai além do espaço de armazenamento e da facilidade de acesso, abordando uma questão mais direcionada para a potencialidade do *Big Data*, que ainda não se consegue utilizar: a reutilização dos dados digitais enquanto recursos econômicos, sociais, culturais, educacionais, científicos, dentre outros campos.

É bastante discutida na área de Organização do Conhecimento a hierarquia Data-Information-Knowledge-Wisdom (DIKW) (Ackoff, 1989). Modelos que incluem hierarquias de níveis crescentes de semântica também são utilizados na Ciência da Computação e na Ontologia. Ao propor a visão da Web Semântica, Tim Berners-Lee (2000) sugere um modelo em camadas, no qual linguagens de crescente expressividade semântica são construídas umas tendo como base outras.

Na Ontologia Formal, é conhecida a proposta de Guarino (2009, p. 6), que estabelece um “nível ontológico”, que seria o nível semântico, no modelo das camadas que constituem as linguagens de representação do conhecimento.

Ao discutir a General Definition of Information (GDI), Floridi (2019) vincula dados, que seriam os elementos básicos da percepção, com informação: informação seria igual a dado + semântica. A definição, aparentemente simples, abre outras perspectivas; embora Floridi não seja o único autor a fazê-lo, sua definição conecta dados, o elemento primário da percepção, com informação. A noção de semântica remete a compreensão, sentido, em especial, no contexto da comunicação humana.

Atualmente, a discussão sobre dados ocupa destaque na Organização do Conhecimento, para além das abordagens de mero insumo para informação e conhecimento. Vários autores destacam duas colocações que são centrais para a análise de dados: o carácter simbólico e intencional dos dados enquanto produto social. Segundo Ackoff (1989, p. 3) “Data are symbols that represent properties of objects, events and their environments”; e “Data are social artefacts” (Ibekwe-Sanjuan; Bowker, 2017, p. 195). Hjørland (2018, p. 9, 25), analisando o *Big Data*, reafirma ambas as colocações e sugere a seguinte definição: “data are concrete instantiations of symbolic representations of descriptive propositions, informed by empirical

observation, about the quantitative and qualitative properties of real-world phenomena"; e "Data are always produced for some purposes and perspectives".

No mesmo texto, Hjørland (2018) vai mais adiante na sua análise, afirmando que "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".

Com base nas definições anteriores propostas Ackoff (1989), Ibekwe-Sanjuan e Bowker (2017) e Hjørland (2018) quanto ao caráter dos dados, podemos sintetizá-las como se segue.

1) **Dados** são ***representações*** de fenômenos ou entidades;
2) Por serem representações, por serem simbólicos, envolvem uma **intenção** de transferir intersubjetivamente conteúdos; estas transferências se ou dão através da Linguagem ou através de artefatos, registros, dados; dados são, portanto, criados e consumidos com determinada **intencionalidade, dados são um produto social**.

Desta forma, nas sociedades humanas, o desenvolvimento das mais diversas atividades sempre deixa traços, ou dados, intencionais ou não, destas atividades (quem exerce a atividade pode não ter a intenção de deixar traços, mas quem coleta estes traços, sim). Dados são, portanto, um produto social. Enquanto representações sígnicas de fenômenos sociais ou físicos de interesse social, dados devem necessariamente ser externalizados em *artefatos, registros*. Os dados são do tipo índices dos fenômenos ou entidades que representam e têm uma relação causal com estes: "índice é um signo que se refere ao objeto denotado em virtude de ser diretamente afetado por esse objeto" (Coelho Neto, 1980, p. 58).

4 TEORIA DOS NÍVEIS INTEGRATIVOS

Compreender a realidade por meio de diferentes níveis hierárquicos distintos é uma concepção antiga, com registro na Filosofia grega antiga. A ideia básica que se coloca é "characterized by the principium rationis sufflcientis which affirms that nothing occurs in the world that does not have its ground in something else". (Hartmann, 1952, p. 69).

Com base nesse pensamento, deu-se origem à Teoria dos Níveis Integrativos, que tem sido aplicada desde a metade do século passado até os dias atuais por vários autores, dentre os quais destacam-se Nicolai Hartmann (1952), pela sua análise detalhada da proposta; e James Feibleman (1954) por sistematizar uma série de doze leis para os níveis, a partir da

contribuição de outros autores (Kleineberg, 2017). A interpretação da teoria, na verdade, difere um pouco de acordo com os autores que a empregam.

Hartmann (1952) propõe dois tipos de hierarquias para compreender a realidade: estratos (*strata*) e camadas (*layers*). As camadas são onde os conceitos são organizados e são constituídas por entidades que mantêm entre si uma relação de constituição hierárquica, onde uma camada de nível hierárquico inferior é constituinte de uma camada de nível hierárquico imediatamente superior, com um conjunto de propriedades diferentes, e que não são herdadas nos outros níveis subordinados. Já os estratos são planos que fornecem visões mais granulares da realidade onde os conceitos se inserem, refletindo uma estruturação dessa realidade, sendo "superimposed one upon the other in all the higher structures" (Hartmann, 1952, p. 51).

Já na acepção de Feibleman (1952), cada nível organiza o nível abaixo acrescentando uma qualidade emergente (que surge quando se muda de nível) e a complexidade dos níveis aumenta de baixo para cima, ou seja, cada nível de cima é mais complexo que o de baixo (como os níveis físico, biológico, social, cultural). Por exemplo, uma pessoa é constituída por células, mas também possui determinadas propriedades que não dizem respeito ao nível da célula, como o funcionamento de um órgão. Além disso, os níveis mais altos dependem dos níveis mais baixos, e, de alguma forma, os impactam.

Conforme aponta Guarino (1999), no âmbito de estudos na área da Ontologia Formal, existe uma ordem intrínseca nas categorias que se situam em níveis ontológicos distintos. Guarino apresenta um conjunto de princípios voltados para a definição de critérios de identidade para estabelecer um conjunto de níveis de abstração que possuem características disjuntas de identidade, propondo um conjunto de níveis ontológicos e exemplificando o que seriam seus respectivos particulares (instâncias no mundo real), observando que estes particulares são únicos para cada nível ontológico.

Dahlberg (1982) usa a teoria dos níveis integrativos para um sistema de classificação universal denominado de Information Coding Classification (ICC), o qual conjugava o uso dessa teoria a uma abordagem facetada para detalhar os campos do conhecimento. Embora o recorte desse esquema seja disciplinar, as disciplinas são organizadas de acordo com níveis de complexidade distintos, sendo as entidades do esquema divididas em três grupos ônticos, a saber: (i) estrutura e matéria; (ii) seres vivos e (iii) produtos humanos (artefatos). Cada um

destes grupos é subdividido em três áreas gerais de entidades, denominadas também de "categorias do ser", permitindo uma visão hierárquica em dois níveis.

Gnoli (2008), por sua vez, adapta a teoria dos níveis integrativos, inspirada em Hartmann (1952), para elaborar uma proposta de sistema classificatório para uso interdisciplinar – Integrative Level Classification[1] (ILC) –, com o foco em recortar a realidade a partir de fenômenos e não em disciplinas. Gnoli propõe, sintetizando propostas de outros autores, a existência de seis estratos (forma, matéria, vida, mente, sociedade e cultura), os quais são formados por uma série de camadas. Segundo a proposta de Gnoli (2008) dados estariam no estrato espírito objetivado, como produtos sociais, criações humanas, artefatos.

Embora a teoria dos níveis integrativos tenha sido largamente utilizada, existem algumas críticas a ela, em especial pelo fato de ser difícil conceber níveis que possam abrigar em um corte preciso as camadas da realidade. As diferentes propostas de níveis parecem ser indício dessa questão.

Eronen (2015) aponta que a natureza é muito complexa para ser enquadrada em um conjunto monolítico, uniforme de níveis hierárquicos, que se aplica a toda a realidade. Propõe como alternativa a adoção de uma abordagem deflacionária, onde em vez de níveis de organização que sejam aplicáveis a tudo, se pense em níveis dedicados a casos específicos: "It is much more plausible that levels are more local and do not extend horizontally across nature. For example, there is a certain hierarchy of levels in a human brain, and a different one in a glacier." (Eronen, 2015, p. 2).

5 RESULTADOS

Esta seção aborda os resultados separados em duas temáticas. A primeira é a conceituação e aplicação dos conceitos de **expressividade semântica de dados digitais** e **saídas** de um sistema de informações. Já a segunda temática ilustra os tipos de agregados de dados organizados em níveis crescentes de expressividade semântica, apresentada no Quadro 1 e seus exemplos e desdobramentos nas Figuras 1 e 2.

5.1 Expressividade semântica de dados digitais e saídas de um sistema de informação

[1] Para maiores detalhes, consultar: http://www.iskoi.org/ilc/book/.

Pode se dizer que expressividade semântica é uma noção usada para designar o quão acuradamente uma representação (mais especificamente uma linguagem de representação) expressa um fenômeno da realidade. A noção é também usada na Organização do Conhecimento; foi cunhada originalmente por Souza, Tudhope e Almeida (2012) para classificar Sistemas de Organização do Conhecimento com crescente capacidade de representar um dado domínio da realidade.

Expressividade semântica poderia também ser chamada de representação de um fenômeno ou fato isolado, representação de uma entidade ou fenômeno complexo, em um domínio da realidade, com suas entidades e relacionamentos, conforme os diferentes níveis; ou o processamento destes agregados de dados gerando respostas conforme o Eixo do Processamento.

Como foi visto na seção 3, dados são representações de aspectos dos mais diversos fenômenos da realidade, criados intencionalmente para permitir agir indiretamente ou de forma mediada, sobre esta realidade. Ao se representar a realidade através de dados, através de Sistemas de Organização do Conhecimento (SOC) e ao se processar estes dados através das tecnologias de Ciência de Dados, gera-se subsídios para agir sobre a realidade.

A partir disso, especulou-se a existência de níveis mais básicos de representação dentro de um ambiente computacional, como os dados dentro de um arquivo segundo um formato, ou um arquivo segundo um formato ou, em um nível ainda mais básico, *bits* e *bytes* formando em uma mídia. Decidiu-se assim, iniciar a análise a partir de dados textuais, que chamamos de Nível 1. Os níveis abaixo deste serão objeto de futuras pesquisas.

A partir daí estes conteúdos textuais se segmentam e agregam com outros conteúdos (possivelmente metadados) formando sistemas de dados, com crescentes níveis de expressividade semântica, fazendo emergir novas categorias semânticas e assim, formando novos níveis.

Os níveis de agregados integrados ou sistemas de dados implicam em uma crescente expressividade semântica. São expressos por novas categorias de agregados de dados, como: colunas de uma tabela, unidade de dados ou *datum* (Hjørland, 2018) – Nível 3; registros de uma tabela ou base de dados ou conceitos/termos (Dahlberg, 1978) como conjunto de propriedades verdadeiras verificadas e consensadas por uma comunidade acerca de um objeto de um determinado domínio – Nível 4; tabelas interrelacionadas segundo um esquema conceitual ou modelo conceitual implícito – Nível 5; ontologias computacionais, agregando

dados – instâncias - e modelos conceituais explícitos destes dados (hierarquia de classes-subclasses, propriedades de objeto e de dados, axiomas) - Nível 6.

Ao se agregarem de forma integrada formando sistemas de dados, dados ganham expressividade semântica, potencializando a emergência de novas categorias de significados, como por exemplo, de conteúdos isolados descontextualizados para uma coluna de uma tabela ou *datum*, ou seja, o valor de uma propriedade de uma dada entidade; ou para várias colunas formando uma tabela completa que representa uma entidade ou fenômeno; para a representação de várias entidades interligadas da realidade.

Os conceitos utilizados na proposta têm as definições sugeridas a seguir.

- Expressividade semântica de agregados de dados digitais: precisão na representação da realidade, informação potencial, potencial informativo da Saída do processamento de um conjunto de dados, oferecida imediatamente a um usuário final, possibilitando-lhe ação imediata ou tomada de decisão; e

- Saída: produto do processamento de um sistema de informação oferecida a um usuário final.

5.2 Agregações de dados em níveis de expressividade semântica e Saídas dos sistemas de informação

Esta subseção apresenta a proposta. Após uma explicação inicial, um Quadro e duas Figuras a ilustram.

Todo modelo genérico de sistema computacional tem dois componentes: dados e programas que os processam. Dados se agregam crescentemente em diferentes níveis, potencializando a geração de significados segundo dois eixos, o Eixo da complexidade das representações (assinalado no Quadro 1 pelos níveis 1, 2, 3, 4, 5 e 6), com agregados de dados representando fenômenos cada vez mais complexos; e o Eixo do processamento destes dados agregados em unidades de representação cada vez mais complexas, gerando significados deste processamento.

O Quadro 1 inicia-se com o nível 0, o nível da Concepção dos dados**.** Neste nível não existem dados realmente, mas somente a concepção, o projeto, um modelo conceitu*a*l dos fenômenos ou coisas que serão representadas pelos dados. No nível mais elevado do quadro 1, o nível 6, os modelos conceituais são incorporados aos dados, nas ontologias computacionais.

A primeira grande mudança de nível no quadro 1 é a divisão clássica em dados não estruturados e dados estruturados. Dados não estruturados são basicamente dados textuais, que são inteligíveis somente por pessoas. Dados estruturados são acompanhados de contexto, representando desde um fenômeno ou fato unitário, até, à medida que se agregam em representações mais complexas, situações, descrições, entidades. Por serem acompanhadas de contexto, são processados por máquina, gerando significados como resposta a esse processamento.

Dados não estruturados textuais passam por diversas etapas de processamento para se tornarem eventualmente dados estruturados. A análise léxica identifica unidades linguísticas discretas em um texto, palavras ou "*tokens*". No Quadro 1 proposto, o resultado da análise léxica de um arquivo, no nível 1, gera dados para o nível 2, na forma de um arquivo de dados textuais com palavras identificadas, uma lista de palavras ou palavras separadas por um delimitador. Igualmente, um arquivo como os descritos para o nível 2 pode ser processado usando técnicas de reconhecimento de entidades nomeadas e a mineração de textos para identificar nomes de entidades de interesse para uma dada aplicação, gerando dados estruturados, contextualizados, como no nível 3.

Dados não têm necessariamente que passar por todos os níveis do Quadro 1. Podem já ser criados segundo qualquer nível de agregação. Exemplos são um arquivo de dados textuais, uma tripla RDF, uma tabela, um conjunto de tabelas formando um esquema de um banco de dados, uma ontologia computacional em OWL.

A partir de determinado nível significados emergem segundo dois eixos – Eixo da Complexidade das Representações – por novas formas, mais complexas, de representar mais acuradamente a fenômenos da realidade, e – Eixo do Processamento – pelo processamento dos fenômenos representadas pelos dados, segundo funções estatísticas como média, média ponderada, desvio padrão, mediana, correlações entre variáveis, etc., largamente utilizadas na Ciência de Dados (Shah, 2020). Estas funções atribuem significados específicos ao resultado deste processamento sobre o comportamento e a dinâmica dos conjuntos de fenômenos analisados. O primeiro eixo de evolução corresponde à Modelagem Conceitual (coisas); o segundo, à Modelagem Estatística (conjuntos ou populações de coisas).

A seguir a proposta é ilustrada. O Quadro 1 ilustra os diferentes níveis de agregação dos dados. Para maior clareza, os exemplos de tipos de dados em cada um dos níveis foram

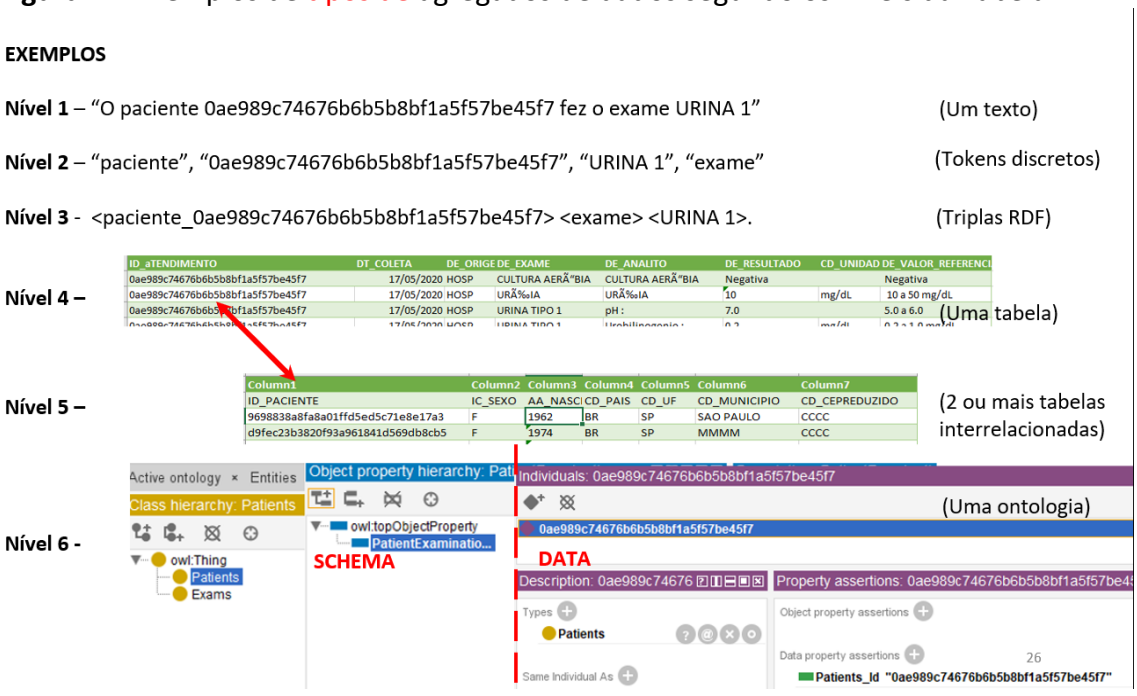
separados na Figura 1. Por sua vez, a Figura 2 mostras as possíveis saídas resultantes do processamento dos agregados de dados dos diferentes níveis [2].

Tabela 1 –Níveis de agregação de dados X expressividade semântica.

		Representação	Processamento	DIMENSÃO TEMPO
Concepção dos dados	**Nivel 0**	**Modelagem conceitual**	**Continuantes**	**Ocorrentes**
			Processamento, consulta	Modelagem estatística
Dados digitais	**Nivel 1**	Dados textuais	PLN, REN	
Dados não estruturados	**Nivell 2**	Tokens discretos, quase-signos (NÖTH, 2002), "data point" (SHAH 2020, p. 16), delimitados mas decontextualizados	MIN. TEXTOS	
Dados estruturados	**Nivel 3**	Dados contextualizados, um estado de coisas (JANSEN, 2008, 188), um "datum" Hjørland (2018): triplas entidade, atributo, valor, a representação de um fato ou fenômeno isolado	SPARQL	Ciên. Dados Dados em tempo real, um Painel
	Nivel 4	Agregação de triplas referenciando uma única entidade ou fenômeno	SQL	
	Nivel 5	2 ou more agregações de triplas referenciando 2 ou mais entidades ou fenômenos interrelacionados; esquema implícito	SQL	
	Nivel 6	Dados incluem o modelo conceitual/esquema; esquema explícito	SPARQL	

Fonte: Elaborado pelos autores.

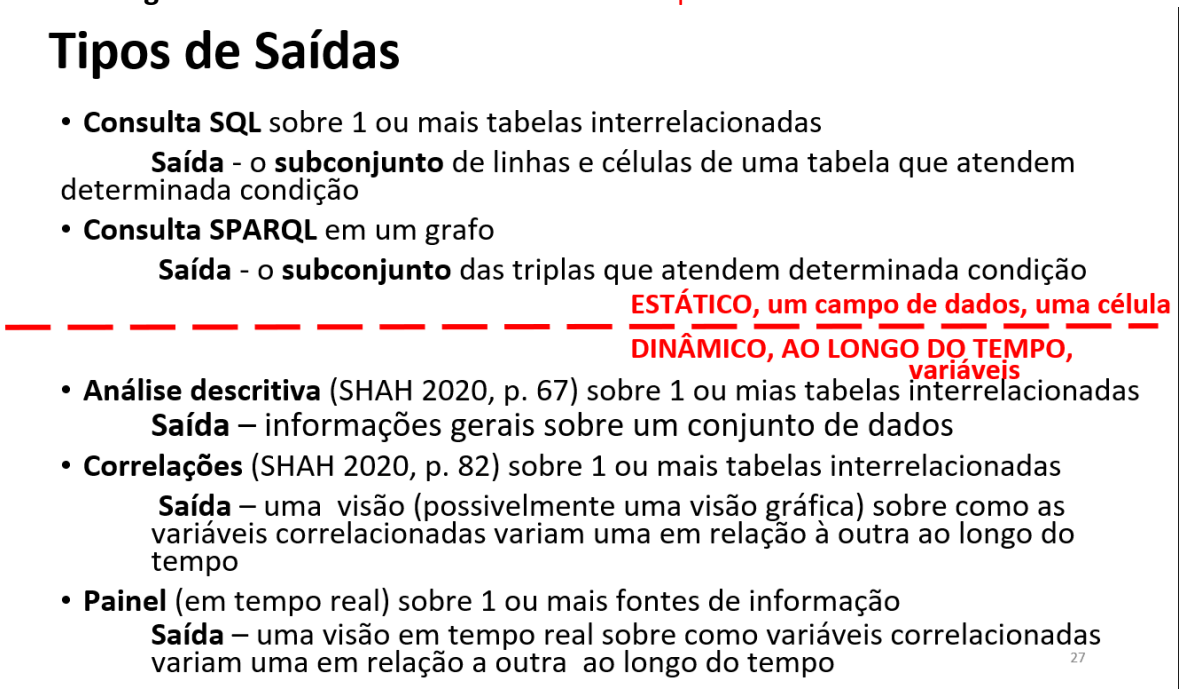
Figura 1 – Exemplos de tipos de agregados de dados segundo os níveis da Tabela 1.



Fonte: Elaborado pelos autores.

[2] Os exemplos citados referem-se a dados obtidos no repositório COVID-19 Data Sharing-BR, mantido pela FAPESP, especificamente ao conjunto Dados COVID Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo.

Figura 2 – Processamento dos diferentes tipos dados e Saídas obtidas



Fonte: Elaborado pelos autores.

A representação do conhecimento tem avançado significativamente a partir da proposta da Web Semântica, a Web de Dados (Berners-Lee, 2000). A classificação das 5 estrelas para a abertura e independência dos dados[3] tem seu nível maior de abertura de dados representados como triplas RDF interligadas, como Dados Abertos Interligados. As triplas RDF tornam cada "*datum*" (Hjørland, 2018) independente de sistemas e aberto. Por outro lado, quando se fala em *Big Data*, os exemplos dados e as mais conhecidas ferramentas para processá-los consideram dados representados como tabelas.

Percebe-se uma convergência cada vez maior entre o *Big Data* e as tecnologias da Web Semântica como dados abertos, grafos de conhecimento e ontologias. É cada vez mais comum que dados, o *Big Data*, sejam gerados como *linked data* e grafos (Debattista et al., 2015). Diante do desenvolvimento de ferramentas de conversão e metodologias de mapeamento, a dicotomia entre dados representados como tabelas e como grafos vai diminuindo e se tornando menos relevante (Hert; Reif; Gall, 2011).

Os Princípios FAIR[4] para dados científicos e tecnologias como grafos de conhecimento e ferramentas de visualização destes apontam novas perspectivas para o tratamento de dados. Os Princípios FAIR agregam aos dados um conjunto de metadados que visam facilitar

[3] Ver https://5stardata.info
[4] Ver https://www.go-fair.org/fair-principles/

sua descoberta (*Findability*), seu acesso (*Accessibility*), sua interoperabilidade (*Interoperability*) e seu reuso (*Reuse*), o objetivo final dos dados abertos. Estes princípios podem e devem ser aplicados a dados não científicos. Com representações como dados interligados, grafos de conhecimento e ontologias, dados ganham ricos contextos, incorporando metadados de proveniência, paradados (HUVILA, 2013).

Juntamente com o *Big Data*, dados em RDF e grafos de conhecimento, colocam novos desafios para a representação do conhecimento. Novos tipos de Saídas, como painéis, agregam dados em linha de diferentes fontes simultaneamente; juntamente com as tecnologias e metodologias de visualização (Hu; Chen, 2021) apontam para novos mecanismos de representação e *síntese* do conhecimento, que podem ser essenciais para extrair "*insights*" de grandes conjuntos de dados - as Saídas, mais complexas - geradas pelo *Big Data*.

6 CONSIDERAÇÕES FINAIS

No contexto do "Big Data", dados, à medida que se agregam em sistemas que representam coisas cada vez mais complexas e em conjuntos numericamente maiores, aumentam sua complexidade semântica, ou seja, seu potencial informativo, ao serem processados.

As ontologias computacionais representam uma síntese das tendências apontadas neste artigo. Juntam, em um único artefato, modelos conceituais (esquemas), representação de fenômenos ou coisas e dados (suas instâncias). O modelo geral delineado aqui integra também, em um ambiente de *Big Data,* dados enquanto representações e seu processamento, a representação sob a forma de dados, ou modelagem conceitual (Organização do conhecimento), e o processamento destas representações, a modelagem estatística (Ciência de Dados), que se utiliza da estatística aplicada, para extrair *insights* de dados quantitativos e qualitativos.

A análise aqui apresentada se constitui em resultados parciais, que deverão ser discutidos e testados para verificar se são uma aproximação consistente da Organização do Conhecimento em direção à Ciência de Dados e *Big Data*.

REFERÊNCIAS

ACKOFF, R. From Data to Wisdom. **Journal of Applied Systems Analysis***,* v. 16, p. 3-9, 1989.

BARRETO, Aldo de Albuquerque. Uma quase história da ciência da informação. **DataGramaZero** - Revista de Ciência da Informação, v. 9, n. 2, p. 1-15, 2008. Disponível em: http://ridi.ibict.br/handle/123456789/162. Acesso em: 21 jun. 2024.

BERNERS-LEE, T. **Semantic Web - XML2000***, 2000*. Disponível em: http://www.w3.org/2000/Talks/1206-xml2k-tbl. Acesso em: 21 jun. 2024.

COELHO NETTO, J. T. **Semiótica, Informação e Comunicação**. São Paulo: Ed. Perspectiva, 1980.

COGNIZANT. Making Sense of Big Data in the petabyte age. **Cognizant**, v. 20, n. 20, Insight, 2011.

DAHLBERG, I. ICC - Information Coding Classification - principles, structure and application possibilities. **International Classification**, v. 9, n.2, p. 87-93, 1982.

DAHLBERG, I. Teoria do conceito. **Ciência da Informação***.* v. 7, n. 2, p. 101-107, 1978. Disponível em: https://doi.org/10.18225/ci.inf.v7i2.115. Acesso em: 05 jul. 2024.

DEBATTISTA, J. et al. Linked'Big'Data: towards a manifold increase in big data value and veracity. In: **2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)**. IEEE, 2015. p. 92-98.

ERONEN, M. Levels of Organization: A Deflationary Account. **Biology & Philosophy***,* v. 30, p. 39-58, 2015. Disponível em: https://markuseronen.com/wp-content/uploads/2017/08/Eronen2015-BP-web-version.pdf. Acesso em: 21 jun. 2024.

FEIBLEMAN, J. K. Theory of Integrative Levels. **The British Journal for the Philosophy of Science**, v5, n.17, p. 59-66, 1954. Disponível em: https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/V.17.59. Acesso em: 21 jun. 2024.

FLORIDI, L. Semantic conceptions of information. In: ZALTA, E.N. (Ed.). **The stanford Encyclopedia of Philosophy**. 2019. Disponível em: http://plato.stanford.edqarchivesqsum2019/entries/information-semantic/. Acesso em: 21 jun. 2024.

FOSKETT, D. J. **Classification and integrative levels**. The Sayers, 1961. Memorial Volume: Essays in Librarianship in Memory of William Charles Berwick Sayers.

FUNDAÇÃO OSWALDO CRUZ. **Plataforma de Ciência de Dados aplicada à Saúde - PCDaS** [Internet]. 2019. Disponível em: https://pcdas.icict.fiocruz.br/sobre-nos/. Acesso em: 9 jul. 2024.

GNOLI, C. Categories and facets in integrative levels. **Axiomathes,** v. 18, n.2, p.177-192, 2008. Disponível em: https://www.gnoli.eu/gnoli2008.pdf. Acesso em: 21 jun. 2024.

GUARINO, N. **The ontological level: Revisiting 30 years of knowledge representation**. Berlin: Springer, 2009. Disponível em: http://telematika.kstu.kg/server/books/ger/conceptualmodel/4.pdf. Acesso em: 21 jun. 2024.

GUARINO, N. The Role of Identity Conditions in Ontology Design. **Proceedings of IJCAI-99 Workshop on Ontologies and problem-solving methods: lessons learned and future trends**. Stockholm, IJCAI, Inc., 1999. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4b93a96fbfb61fb7bf7f52bd535f399f280a21b2. Acesso em: 21 jun. 2024.

HARE, Maddie. Towards Wisdom: Knowledge Management and the Ethical Use of Big Data in Organizational Decision-Making. **Dalhousie Journal of Interdisciplinary Management**, v. 17, 2023.

HARTMANN, N. **New ways of ontology**. Henry Regnery Company, Chicago, Illinois, 1952.

HERT, M.; REIF, G.; GALL, H. C. A comparison of RDB-to-RDF mapping languages. In: **Proceedings of the 7th international conference on semantic systems**. 2011. p. 25-32.

HJØRLAND, B. Data with big data and database semantics. **IEKO, ISKO Encyclopedia of Knowledge Organization**. ISKO, 2018.

HU, A.; CHEN, H. Data visualization analysis of knowledge graph application. In: **2021 2nd International Conference on Artificial Intelligence and Information Systems**. 2021. p. 1-10.

HUVILA, Isto. The unbearable complexity of documenting intellectual processes: Paradata and virtual cultural heritage visualisation. **Human IT: Journal for Information Technology Studies as a Human Science**, v. 12, n. 1, 2013.

IBEKWE-SANJUAN, F.; BOWKER, G. C. Implications of Big Data for Knowledge Organization. **Knowledge Organization**, v. 44, n. 3, p. 187-198, 2017.

KLEINEBERG, Michael. Integrative Levels. **Knowledge Organization**, v. 44, p. 349-379, 2017.

LATOUR, B. Redes que a razão desconhece: laboratórios, bibliotecas, coleções. In: BARATIN, M; JACOB, C. **O poder das bibliotecas**. 2.ed. Rio de Janeiro: Ed. UFRJ, 2006.

MEJIAS, U. A.; COULDRY, N. Datafication. **Internet Policy Review**, v. 8, n. 4, 2019. Disponível em: https://policyreview.info/concepts/datafication. Acesso em: 21 jun. 2024.

SHAH, C. **Hands-on Introduction to Data Science**. Cambridge: Cambridge University Press., 2020.

SOUZA, R. R.; TUDHOPE, D.; ALMEIDA, M. B. Towards a taxonomy of KOS: dimensions for classifying Knowledge Organization Systems. **Knowledge Organization**, v. 39, n. 3, p. 179-192, 2012.