Arquivologia e Ciência da Informação na Era do Big Data: Perspectivas de Pesquisa e Atuação Profissional em Arquivos Digitais

Archival Science and Information Science in the Age of the Big Data: Research Perspectives and Professional Performance in Digital Archives

Jonas Ferrigolo Melo

Universidade Federal do Rio Grande do Sul
jonasferrigolo@gmail.com

Moisés Rockembach

Universidade Federal do Rio Grande do Sul
moises.rockembach@gmail.com

Resumo

Este artigo procura identificar os conceitos e delimitações no campo de investigação dos grandes conjuntos de dados, ou *Big Data*, área emergente que envolve principalmente as tecnologias de informação e comunicação, mas também possui ligações interdisciplinares, abordando aqui a conexão com os profissionais da informação. Demonstra a produção científica nos temas *Big Data* e Cientista de Dados, procurando verificar os principais artigos publicados. Explora a abordagem do dataism, como transformação da ação social em dados online, permitindo o rastreamento em tempo real e análise preditiva de determinadas ações. Traz, por fim, reflexões acerca do *Big Data* como importante elemento de investigação científica e objeto de trabalho para os profissionais da informação.

Palavras-chave: Arquivologia. Ciência da Informação. Ciência de dados. *Big Data*. Dataficação.

Abstract

This paper aims to identify the concepts and delimitations in the field of investigation of Big Data, an emerging area that mainly involves information and communication technologies, but also has interdisciplinary connections, approaching here the connection with information professionals. It demonstrates the scientific production in the subjects of Big Data and Data Scientist, looking for to verify the main published papers. It explores the approach of dataism, how to transform social action into online data, allowing for real-time tracking and predictive analysis of certain actions. Finally, it has reflections on Big Data as an important element of scientific research and work object for information professionals.

Keywords: *Archival Science. Information Science. Data science. Big Data. Datafication.*

1. Introdução

Na era dos grandes conjuntos de dados, ou do "*Big Data*", no qual nos encontramos, a estrutura da informação torna-se particularmente complexa, tanto pelo volume de dados gerados, como pelas características de arquitetura distribuída, pois saímos de um contexto local para um ambiente de sistemas distribuídos.

Atualmente, a organização e recuperação de dados tem se tornado um problema a ser enfrentado, dada a grande quantidade de informações que temos que administrar. A Sociedade vem passando, especialmente na última década, uma grande transformação em relação a quantidade, variedade e volume de dados e, por consequência, demandando conhecimentos específicos para algumas áreas do conhecimento.

O que nos interessa para a reflexão que ora se propõe é que a Era da Informação tem, cada vez mais, focado no indivíduo e tem deixado de lado o mercado de massas: os serviços saíram da massificação para a customização, transferindo cada vez mais poder à personalização (Jenkins, 2009).

Consequentemente, para que a personalização possa acontecer, há a necessidade de que a demanda seja conhecida. Em função disto, o tratamento e o uso da informação tem se modificado nos últimos anos em consequência do surgimento destas novas necessidades sociais, econômicas, comerciais e até mesmo tecnológicas. Estas modificações estão promovendo uma alteração exponencial de paradigma, desta vez voltado para os dados, que poderá ser tão relevante quando da Revolução da Imprensa, de Gutenberg. Desta forma podemos colocar que "[...] a era dos grandes conjuntos de dados está apenas no início, mas já é importante começarmos a questionar os pressupostos, valores e vieses dessa nova onda de pesquisa." (Boyd; Crawford, 2012: 24-25, tradução nossa).[1]

A crescente utilização de meios de comunicação com alto grau de mobilidade e o uso cada vez maior da Internet definem outros espaços e demarcam novas fronteiras para a sociedade contemporânea (Ribeiro, 2014: 97). Neste cenário repleto de dados e informações em diferentes fontes, suportes e formatos, vem emergindo um novo nicho de estudo: o *Big Data*.

Mayer-Schönberger e Cukier (2013) dizem que o *Big Data* tem-se tornado uma das principais fontes de renda e informações de muitas organizações. Demandas organizacionais e governamentais, dada a evolução tecnológica, por si só, em seu quotidiano, exige e armazena uma quantidade imensa de informações de seus usuários, afinal, estamos em uma sociedade marcada pela produção, circulação, armazenamento e controle de uma quantidade massiva de dados.

Para além de ser um resultado da explosão informacional vivida na atualidade, o *Big Data* também tem um grande desafio sob a necessidade de proporcionar a recuperação e acesso a toda essa informação. Percebe-se a necessidade de uma mudança na forma de agir do profissional da informação, não em relação ao seu objeto de trabalho, que segue sendo a informação em qualquer de seus formatos, mas em relação a quantidade e a velocidade que se faz necessária para seu gerenciamento, armazenamento e, especialmente, sua recuperação. Essa mudança de paradigma trouxe novas exigências para os profissionais que atuarão com o gerenciamento de dados. Não se trata

1 "The era of Big Data has only just begun, but it is already important that we start questioning the assumptions, values, and biases of this new wave of research". (Boyd; Crawford, 2012: 24-25).

apenas de algo a mais na formação do profissional da informação, mas em uma mudança exponencial na forma como esses profissionais lidam com dados, informações, documentos e tecnologia.

Boyd e Crawford (2012) dizem que existe um profundo impulso governamental e industrial no sentido de coletar e extrair o máximo de dados possível, seja a partir de informações que possibilitem a efetivação de uma publicidade mais direcionada, de design de produtos, planejamento de tráfego ou policiamento. Existem implicações sérias e complexas para a operacionalização e uso de grande quantidade de dados, considerando que serão ferramentas muito utilizadas para futuras agendas de pesquisa. Neste sentido, o mercado de trabalho passará a exigir características e conhecimentos específicos para quem for atuar com grandes dados.

Este artigo busca conhecer a produção científica sobre o *Big Data* e o profissional de Ciência de Dados, com a intenção de refletir sobre como os profissionais da informação poderão se encaixar neste novo nicho mercadológico da informação e quais serão as exigências para este profissional. Antes disso, acreditamos que se faz necessária uma rápida explanação sobre o *Big Data*, seus conceitos e características.

2. Conceitos e delimitações acerca do *Big Data*

Boyd e Crawford (2012: 2) dizem que o termo *Big Data* vem sendo utilizado como forma de identificar apenas os grandes dados, porém entendem que *Big Data* tem relação mais ampla: para além de grandes dados, o termo refere-se à capacidade de pesquisar, agregar e fazer referências cruzadas destes conjuntos de grandes dados. Na tentativa de conceituar o termo *Big Data*, as autoras definem como um fenómeno cultural, tecnológico e académico que repousa na interação de Tecnologia, sendo a capacidade de utilizar a computação para reunir, analisar, vincular e comparar se utilizando de algoritmos; Análise, como a possibilidade de identificar padrões a fim de reivindicações econômicas, sociais, técnicas e legais; e Mitologia, que se refere a uma crença generalizada de que grandes conjuntos de dados oferecem uma forma superior de inteligência e conhecimento que pode gerar percepções que antes eram impossíveis, com tamanha veracidade, objetividade e precisão. (Boyd; Crawford, 2012: 3).

A *TechAmerica Foundation* (2012) traz uma definição mais clara, dizendo que "*Big Data* é um termo que descreve grandes volumes de dados variáveis, complexos e de alta velocidade que requerem técnicas e tecnologias avançadas para permitir a captura, armazenamento, distribuição, gerenciamento e análise da informação." (*TechAmerica Foundation*, 2012, tradução nossa).[2]

Já o glossário Gartner define da seguinte forma: "*Big Data* é um grande volume de informações, alta velocidade e alta variedade de ativos de informação que demandam formas inovadoras e econômicas de processamento de informações para melhor percepção e tomada de decisão." (Gartner, 2018, tradução nossa).[3]

2 "Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the informacion." (TechAmerica Foundation's…, 2012).

3 "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". (Gartner, It Glossary, n.d.).

Autores como Dumbill (2012); Nerurkar; Wadephul; Wiegerling (2016) e Zikopoulos et al (2011), dizem que o *Big Data* está apoiado em três fatores de sustentação conhecidos como os três V's do *Big Data*: Volume, Velocidade e Variedade:

O "volume" está ligado a grande quantidade de dados e informações que nos cercam no cotidiano. A sociedade da informação na era digital, considerando a multiplicidade de dispositivos e sua capacidade de interação em rede, está fortemente marcada pela produção, circulação, armazenamento e controle de uma quantidade expressiva de dados. Um dos benefícios obtidos em relação ao *Big Data* diz respeito à possibilidade de processar grandes quantidades de informações e, ao mesmo tempo, é um dos seus principais desafios, considerando que o volume de dados pode ser maior do que a infraestrutura de bancos de dados relacionais convencionais podem suportar. Ribeiro (2014: 97) corrobora dizendo que multiplicidade de informações que circulam na internet, originadas pelas diferentes fontes, ocasionam a sobrecarga de informações que são disponíveis à sociedade, e que apenas 1% destes dados são efetivamente analisados. Em função disso, o autor deduz que há um grande campo de atuação aos profissionais que atuam com a gestão da informação.

A "velocidade" do processamento e da fruição dos dados é um dos pilares sustentadores do *Big Data*. A melhoria dos canais de transmissão, com redes em fibra ótica e emissores de sinais de alta capacidade, o uso de satélites, o uso de outras bandas para a telefonia celular, as comunicações em tempo real para controle de processos na internet, os *workflows* científicos com processamento paralelo e *cluster* de processamentos vem possibilitando atingir maior velocidade para troca de dados e informação (Mattoso, 2013 apud Ribeiro, 2014: 100). O uso da Internet, por meio de celulares, computadores portáteis, relógios e tantos outros equipamentos que acessam a *Web*, proporcionou uma grande captura de dados que pudessem ser importantes para o mercado facilitado por meio da internet.

O terceiro pilar de sustentação da teoria do *Big Data* é a "variedade" dos tipos de dados produzidos a partir do uso de dispositivos ligados a *Web*. Cabe salientar que raramente os dados se apresentam em uma forma perfeitamente ordenada e pronta para processamento. Um tema comum em sistemas de *Big Data* é que os dados capturados são diversos e não se encaixam em estruturas relacionais perfeitas. Podem ser texto de redes sociais, dados de imagem e outras informações mais precisas e direcionadas, ou um *feed* bruto em que há necessidade de interpretação dos dados.

Volume, velocidade e variedade são os V's que mais aparecem nas literaturas da área. Porém, cabe destacar que outros autores expõem diferentes V's como os pilares do *Big Data*, a exemplo de Gandomi e Haider (2015) que discorrem sobre Veracidade, Variabilidade e Valor a partir do uso de diferentes fontes:

A "veracidade", segundo os autores, foi o V introduzido pela IBM[4] como o quarto V. Ele representa a falta de confiabilidade inerente a algumas fontes de dados. Por exemplo, os sentimentos dos usuários de mídias sociais são incertos, uma vez que envolvem o julgamento humano. No entanto, eles contêm informações valiosas. A necessidade de lidar com dados imprecisos e incertos seria outra faceta do *Big Data*, que poderia ser explorada com o uso de ferramentas e análises desenvolvidas para

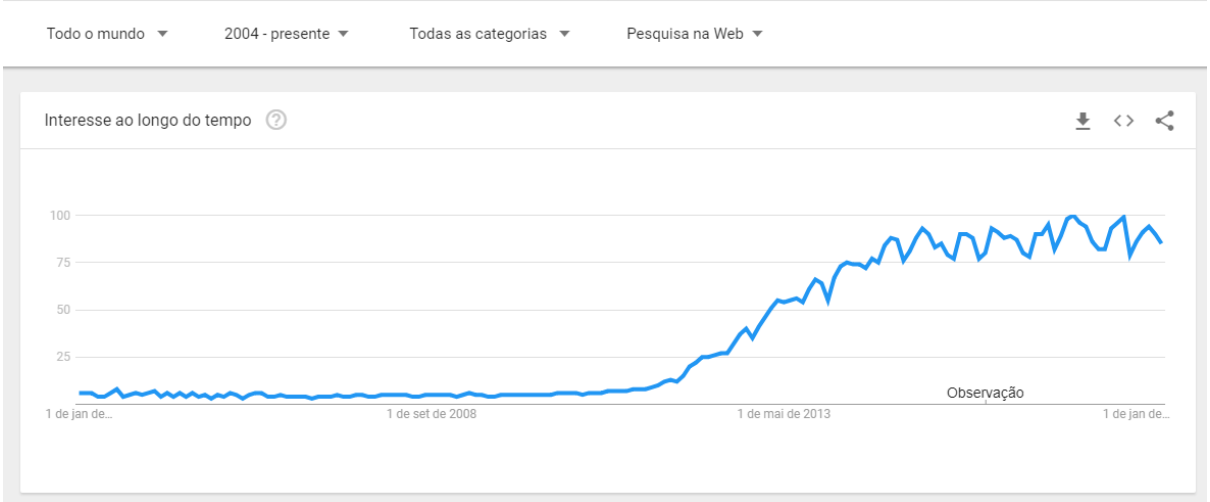
4 IBM (International Business Machines) é uma empresa dos Estados Unidos que fabrica e vende hardware e software, fundada em 1924. https://www.ibm.com

gerenciamento e mineração de dados incertos. A "variabilidade" foi introduzida pela SAS Inc[5] como dimensões adicionais de *Big Data*, considerando que emissão e captura de grandes dados não é constante em termos de volume; variabilidade refere-se à essa variação nas taxas de fluxo de dados. O "valor" foi o termo apresentado pela Oracle[6] como um atributo definidor de *Big Data*. Com base na definição da Oracle, os grandes dados, quando recebidos na forma original, geralmente, têm um valor informacional baixo em relação ao seu volume. No entanto, um alto valor informacional pode ser obtido através da análise desses dados.

3. Interesse e investigação científica sobre *Big Data*

A partir do *Google Trends* percebe-se um crescimento exponencial da busca pelo termo "*Big Data*", de 2004 até os dias atuais, em todo o mundo (Gráfico 1).

Gráfico 1 – Desenvolvimento de buscas no Google sobre Big Data



Fonte: Google Trends, 2018.

Da mesma forma, foram pesquisados os artigos científicos indexados na *Web of Science* - WoS[7], que apresentavam o termo *Big Data* em seus metadados. No período de 1954 até 2018 foram indexados 31.216 artigos. Destes, 19.467 artigos científicos foram indexados nos últimos cinco (05) anos, de 2014 a 2018. Somente em 2017 até os dias atuais foram indexados na WoS 9.223 artigos. Um dado interessante que se pode notar é que de 2004 a 2018 foram 27.865 artigos científicos, portanto de 1954 até 2003 foram indexados apenas 2.906 artigos sobre *Big Data*. Nota-se um crescimento exponencial da produção de trabalho científicos em que *Big Data* está figurando como assunto.

Na busca por textos, autores e referências sobre o cientista de dados, utilizamos os artigos indexados na *Web of Science* (WoS) em que configuravam no item "tópico" os termos *Big Data AND data scientist*. Foram encontrados 485 artigos, indexados na base, desde 1991 até 21 de julho de 2018 (25 anos). Somente nos últimos 10 anos, portanto de 2009 a 2018, foram indexados 439 artigos. Esse dado corrobora para exemplificar o crescimento exponencial sobre o campo de estudos do *Big Data*.

5 SAS Inc. (Statistical Analysis System) é uma organização norte americana, localizada no estado da Carolina do Norte (EUA). Fundada em 1976, é produtora de softwares para Business Intelligence. https://www.sas.com

6 Oracle Corporation é uma organização norte americana, fundada em 1977, especializada no desenvolvimento de hardware, softwares e banco de dados. https://www.oracle.com

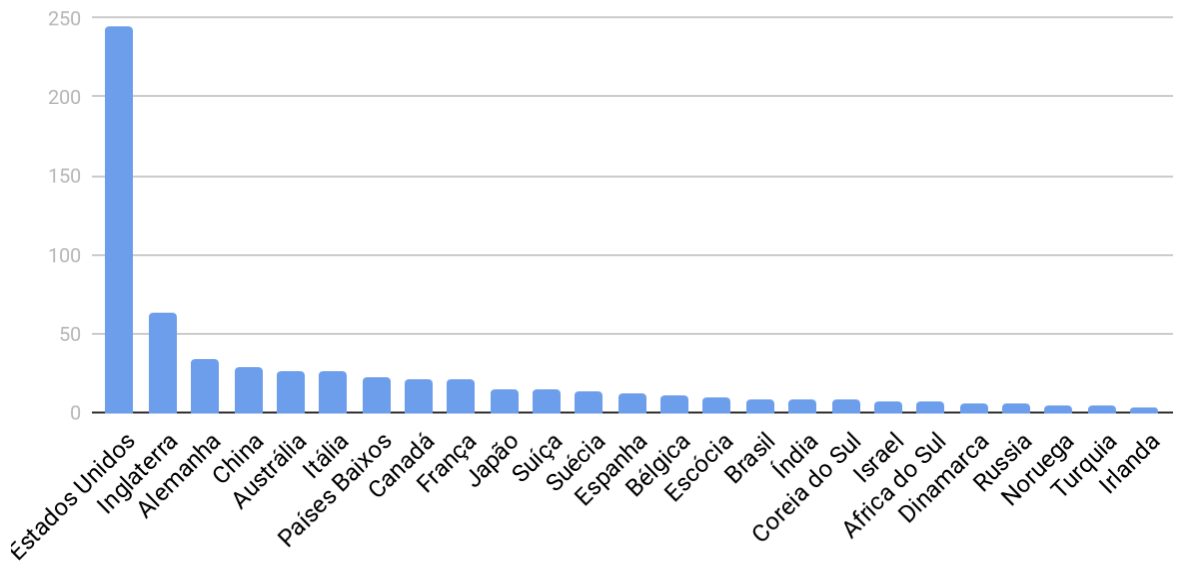
7 A pesquisa na base de dados Web of Science foi realizada em dezembro de 2018.

Tabela 1 – Artigos científicos sobre Big Data e Data Scients indexados na WoS

Período analisado	Quantidade de artigos	Percentual produção por período
de 1991 a 2018 (25 anos)	485 artigos	100%
de 2009 a 2018 (10 anos)	439 artigos	90,52%
de 1991 a 2008 (15 anos)	46 artigos	9,48%

Fonte: Dos autores.

Gráfico 2 – Quantidade de artigos por países



Fonte: Dos autores.

Seguindo a análise dos artigos, foram identificadas 154 Categorias da Web of Science (campo WC) em que os arquivos foram indexados. As dez categorias com mais artigos indexados são: Computer Science Interdisciplinary Applications; Computer Science Theory Methods; Computer Science Information Systems; Ecology; Environmental Sciences; Information Science Library Science; Computer Science Software Engineering; Social Sciences Interdisciplinary; Multidisciplinary Sciences e Management.

Em relação aos autores, verificamos a frequência com que aparecem nos artigos selecionados. Inicialmente, ao utilizar o campo AU da WoS, foram extraídos 1.962 autores, sendo que 1.914 autores publicaram 1 artigo cada; 46 autores publicaram 2 artigos e 2 autores publicaram 3 artigos.

Num segundo momento verificamos os dados do campo AF da WoS, que se refere ao nome completo do autor. Ao analisar os dados percebemos uma discrepância: foram extraídos um total de 1.984

autores, sendo que 1.957 publicaram apenas 1 artigo; 26 autores publicaram 2 artigos e apenas 1 autor publicou 3 artigos, conforme especificado na tabela a seguir:

Tabela 2 – Comparativo da pesquisa dos campos AU e AF da WoS

	Resultado campo AU	**Resultado campo AF**
Total de autores	1962	1984
Autores com 1 artigos publicado	1914	1957
Autores com 2 artigos publicados	46	26
Autores com 3 artigos publicados	2	1

Fonte: Dos autores.

Ao comparar o resultado de ambos os campos, foi constatada a diferença de 22 autores. Neste sentido, percebemos que o uso do campo AU para esta análise pode ocasionar falha no resultado, considerando a quantidade elevada de homônimos para nome dos autores. Um dos casos que cabe destacar é do autor Chen J., que aparecia no campo AU com 3 artigos publicados, mas na verdade se tratavam de 3 autores diferentes, como demonstrado no resultado do campo AF: Chen, Jing; Chen, Jian e Chen, Jie.

Em relação a quantidade de citações de cada artigo estudado, constatamos que há uma diferença exponencial, que poderá ser verificada no quadro a seguir, em que apresentamos os dados dos seis artigos mais citados:

Tabela 3 – Ranking de citações

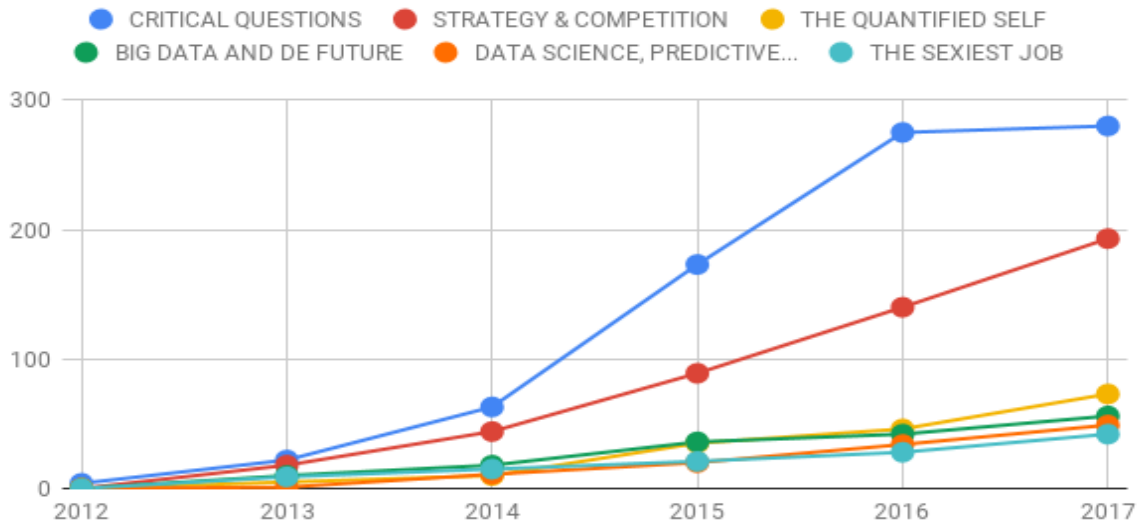
Identificador	***Título***	***Autoria***	***Ano de Publicação***	***Quantidade de citações***
1	***CRITICAL QUESTIONS FOR BIG DATA Provocations for a cultural, technological, and scholarly phenomenon***	***Boyd, Danah; Crawford, Kate***	***2012***	***963***
2	***STRATEGY & COMPETITION Big Data: The Management Revolution***	***McAfee, Andrew; Brynjolfsson, Erik***	***2012***	***586***
3	***THE QUANTIFIED SELF: Fundamental Disruption in Big Data Science and Biological Discovery***	***Swan, Melanie***	***2013***	***200***
4	***Big Data and the future of ecology***	***Hampton, Stephanie E.; Strasser, Carly A.; Tewksbury, Joshua J.;***	***2013***	***187***

		Gram, Wendy K.; Budden, Amber E.; Batcheller, Archer L.; Duke, Clifford S.; Porter, John H.		
5	***Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management***	***Waller, Matthew A.; Fawcett, Stanley E.***	***2013***	***163***
6	***Data Scientist: The Sexiest Job of the 21st Century***	***Davenport, Thomas H.; Patil, D. J.***	***2012***	***139***

Fonte: Dos autores.

Ao analisar a tabela percebe-se que os artigos "Critical Questions For *Big Data*" e "Strategy & Competition" são exponencialmente os mais citados entre os artigos analisados. "The Quantified Self", o terceiro artigo mais citado, está distante em quantidade de citações em relação ao segundo artigo mais citado.

Gráfico 3 – Evolução da quantidade de citações por artigo



Fonte: Dos autores.

No gráfico acima apresentamos o gráfico que mostra a evolução das citações por ano e por artigo. Percebe-se um crescimento constante, a cada ano, na quantidade de citações para todos os artigos analisados.

Atualmente, os dados estão surgindo em muitos formatos e podem ser estruturados, semi-estruturados, não estruturados e até mesmo dados estruturados complexos. Essa grande variedade requer uma abordagem diferente, bem como diferentes técnicas para armazenamento de todos os dados brutos, que por sua vez, requerem diferentes análises ou ferramentas. Um uso comum do processamento de *Big Data* é obter dados e extrair o significado ordenado para consumo por empresas

e outros agentes que, a partir destas informações, conseguem monetizar seus negócios. Essa prática é chamada de "*datafication*" ou dataficação.

4. Reflexões para os profissionais da informação em um mundo de dataficação

A informação é um processo de criação de sentido a partir do uso da tecnologia da informação (Lycett, 2013: 383). A construção de sentido frente a informação refere-se aos processos de organização dos dados a partir do uso da linguagem para identificar e normalizar estes dados, dando-lhes características identificáveis, ou seja, dando-lhes sentido. Essa significação diz respeito às informações que as pessoas geram e o que pode ser interpretado em termos de interesses, emoções e demais rastros deixados a partir de suas interações com as redes online. Cabe destacar que a revolução real não está nas máquinas que calculam os dados, mas nos dados em si e em como os utilizamos.

A forma como são entregues os dados faz com que também sejam consumidos produtos e serviços de forma cada vez mais instrumentalizada, gerando um fluxo de dados que voltam para o provedor (Dumbill, 2012: 7, tradução nossa) que, por sua vez, processa e faz com que retorne informações personalizadas aos usuários. Quantas vezes já fizemos busca no Google, por exemplo, para comprar uma passagem aérea e de forma instantânea todas as nossas redes sociais nos apresentam promoções de passagens aéreas exatamente para aquele destino que pesquisamos? Os resultados podem ir diretamente para um produto, como para as recomendações do Facebook ou em painéis usados para orientar a tomada de decisões. Os dados são processados e retornam de maneira personalizada aos nossos interesses. Este cenário reflete bem como a velocidade age em relação ao *Big Data*: os dados são fornecidos a partir de nossos acessos.

> *O avanço do uso de dispositivos móveis, o uso de sensores industriais e biomédicos, fotos, vídeos, e-mails, redes sociais, além do comércio eletrônico, interações via call centers, dispositivos móveis, dados públicos, imagens médicas e outros dados científicos, câmeras para monitoramento, medidores inteligentes, GPS, aplicativos para troca de mensagens, aplicações que nos ajudam a pegar táxis, outras que nos ajudam na locomoção urbana evitando engarrafamentos, ou ainda no monitoramento de ônibus e até de aviões, são exemplos concretos desta avalanche. (Ribeiro, 2014: 98).*

Os varejistas *online* são capazes de compilar um grande histórico de cliques e interações dos usuários/clientes, não apenas suas vendas finais. Aqueles que são capazes de utilizar rapidamente a informação disponibilizada por meio dos cliques do usuário/cliente em potencial, recomendando compras adicionais, por exemplo, e ganham vantagem competitiva. O uso massivo do *smartphone* aumenta exponencialmente a taxa de entrada de dados, já que os consumidores carregam consigo uma fonte de imagens, mensagens de texto, dados de áudio, trajetos por onde circulam, etc. Não se trata apenas da velocidade com que os dados são recebidos pelo processador, mas especialmente em relação a transmissão destes dados. A importância está na velocidade do *feedback* (Dumbill, 2012: 7, tradução nossa), que segundo Ribeiro (2014: 100) se tornará cada vez mais veloz, considerando o desenvolvimento da tecnologia de processadores, de canais e de *hardware* para armazenamento, fazendo com que o poder da velocidade seja duplicado a cada período de dois anos, proporcionando, também, o desenvolvimento de sistemas de armazenamento e busca por métodos mais ágeis e eficazes, otimizados para a rápida recuperação de informações.

Parte da explicação destes fenómenos pode ser encontrada na gradual normalização da informação como um novo paradigma na Ciência e na Sociedade. *Datification ou* dataficação, segundo Mayer-Schöenberger e Cukier (2013), é a transformação da ação social em dados online tabulados e quantificados, permitindo assim o rastreamento em tempo real e a análise antecipada de determinadas ações, em outras palavras, "refere-se a tornar informação todas as coisas sob o sol - incluindo aquelas que nunca pensamos como informação [...]"[8]; documentar uma ação é "[...] colocá-las em um formato quantificado para que possam ser tabuladas e analisadas"[9] (Mayer-Schöenberger; Cukier, 2013, n/p., tradução nossa).

Agências governamentais e empresas se apropriam da coleta de metadados gerados a partir do uso de sites de redes sociais, plataformas de comunicação, serviços de *streaming*[10] e de e-mails gratuitos, a fim de rastrear informações sobre o comportamento humano. "Agora podemos coletar informações que não podíamos antes, sejam de relacionamento através de telefonemas ou sentimentos revelados por meio de tweets"[11] (Mayer-Schöenberger; Cukier, 2013: s/p., tradução nossa). A forma de distribuição de comunicação personalizada como um meio legítimo de entender e monitorar o comportamento das pessoas está se tornando um princípio de liderança, não apenas entre os avanços tecnológicos, mas, especialmente no mercado e no meio académico. Ambos nichos veem a *datafication* como uma oportunidade de atuação revolucionária para investigar a conduta humana, seja para utilizá-la como dado científico ou para ampliar seus negócios lucrativos. Este fenômeno foi identificado pelo atributo *valor* em *Big Data* e tornou-se um dos pilares balizadores da teoria.

Fazer algo valioso com os dados é a razão para armazená-los. Organizações buscam aumentar seu desempenho a partir da mineração de dados, com o intuito de prever e orientar estratégias futuras e operações do dia-a-dia. *Datafication* fornece a lente para essas previsões: ajuda as empresas a "[...] prever eventos antes que eles aconteçam"[12] (Mayer-Schöenberger; Cukier, 2013: s/p., tradução nossa), permitindo-lhes tomar decisões informadas e lucrativas. Mayer-Schönberger e Cukier (2013) explicam que com o uso da *datafication* pode-se fazer com que palavras se tornem dados, permitindo que as empresas façam pesquisas e as transformem em inferências analíticas de dados. É inegável que a maneira como entregamos e consumimos produtos e serviços está cada vez mais instrumentalizada. O uso de informações pessoais, a partir da *datafication*, constitui um novo tipo de sociedade da informação (Mai, 2016: 193).

A ideia de que os dados são recursos "brutos" esperando para serem processados se encaixa perfeitamente na metáfora de "mineração da vida":

> *Estamos testemunhando o surgimento de um novo tipo de extração de informações que talvez seja melhor caracterizado como "mineração da vida": extrair conhecimento útil dos rastros digitais deixados por pessoas que vivem online uma parte considerável de sua vida. A possibilidade de prever atividades é um caso especial da mineração da vida. (Weerkamp, Rijke, 2012, s/p., tradução nossa).*[13]

8 "It refers to taking information about all things under the sun—including ones we never used to think of as information at all [...]". (Mayer-Schöenberger; Cukier, 2013: s/p.).

9 "[...] put it in a quantified format so it can be tabulated and analyzed." (Mayer-Schöenberger; Cukier, 2013, n/p.).

10 *Streaming* é uma tecnologia que envia informações multimídia através da transferência de dados, utilizando redes de computadores, especialmente a Internet.

11 "We can now collect information that we couldn't before, be it relationships revealed by phone calls or sentiments unveiled through tweets." (Mayer-Schöenberger; Cukier, 2013: s/p.).

12 "One of these is a method called predictive analytics, which is starting to be widely used in business to foresee events before they happen." (Mayer-Schöenberger; Cukier, 2013: s/p.).

13 "We are witnessing the emergence of a new type of time-aware information extraction that is perhaps best characterized as "life mining": extracting useful knowledge from the combined digital trails left behind by people who live a considerable part of their life online. Activity prediction is a special case of life mining." (Weerkamp, Rijke, 2012, s/p.).

Mayer-Schöenberger e Cukier (2013) dizem que a metáfora da mineração da vida está fundamentada em uma lógica peculiar que orienta empresas, pesquisadores e agências estatais na busca por um valor intrínseco aos conjuntos de dados, ainda ocultos. As empresas estão envolvidas em uma corrida para descobrir como capturar e avaliar esse valor. A *datafication* cresceu e tende a crescer cada vez mais, tornando-se um novo paradigma usado para entender o comportamento social. Com o advento da *web* e a proliferação de sites de redes sociais e plataformas de *streaming*, muitos aspetos da vida social foram codificados: amizades, interesses, relacionamentos, conversas casuais, pesquisas, manifestação de gostos e emoções (Van Dijck, 2014: 198). Além disso, as empresas de tecnologia codificaram outros aspetos do dia a dia, como por exemplo, acompanhar a vida de um familiar através de um *followers*; a atividade de contar algo sobre alguém através de um *retweet;* ou de procurar um emprego por meio de uma interface digital; mostrar a música do momento através de um *link*, e assim por diante. À medida que essas ferramentas foram surgindo, a sociedade transferiu e confiou, de certa forma, vários aspetos de suas interações sociais e pessoais aos ambientes da *web* e às empresas de tecnologia, autorizando que seus dados pessoais fossem minerados e que nossas atividades fossem premeditadas.

Van Dijck (2014) chama este novo paradigma sociocientífico de *Dataism*. Este paradigma revela uma crença na quantificação e no potencial de rastrear todos os tipos de comportamento humano e social através de dados online, obtidos por meio de tecnologias de mídias sociais. Considera, também, que estes dados são apresentados como "matéria-prima" e que podem ser analisados e processados em algoritmos com o potencial de prever o futuro comportamento humano (Van Dijck, 2014: 198). Além disso, o *dataism* envolve a confiança nos agentes institucionais que coletam, interpretam e compartilham os dados oriundos das mídias sociais, plataformas de internet e outras tecnologias de comunicação.

O *dataism* acredita que a coleta de dados pode acontecer independente de qualquer estrutura predefinida e a análise dos dados acontece sem, necessariamente, um propósito predefinido - como se os mineradores de dados analisassem essa coleta apenas para acumular conhecimento sobre o comportamento das pessoas (Van Dijck, 2014: 202).

A *datafication* e a mineração da vida são apostas em pressupostos ideológicos, oriundo de uma perspectiva teórica que deu origem a um novo paradigma científico, o *Dataism*. Essas teorias estão enraizadas nas normas sociais vigentes, especialmente quando voltamos nossas relações sociais à *web*. As empresas de tecnologia fornecem serviços aos usuários, que por sua vez fornecem informações pessoais às empresas. Há uma permuta em que a moeda são os metadados, um tipo de ativo invisível: as empresas de mídia social monetizam os metadados ao interpretá-los e vendê-los a anunciantes ou empresas de dados. A *datafication* é a descoberta da forma de minerar os dados que são, aparentemente, não-vulneráveis. A verdade é que, hoje em dia, quase tudo pode ser um dado: desde uma busca na *web*, até o movimento de um *smartphone* no bolso de seu usuário, pois "existe um profundo impulso governamental e industrial em coletar e extrair valor máximo dos dados" (Boyd e Crawford, 2012: 14, tradução nossa).

A partir destes dados, pode-se perceber que o interesse pelo *Big Data* tem sido amplamente aumentado nas áreas acadêmica e empresarial, se tornando evidente a necessidade de preparo de um profissional específico para essa recente demanda de trabalho. Será o profissional de Tecnologia da Informação (TI) capaz de absorver todo este mercado? Ou serão também os Arquivistas, Bibliotecários e Cientistas da Informação, a partir de uma ligação interdisciplinar? Será que os currículos das

academias que formam estes profissionais propensos a assumir o mercado do *Big Data* estão adaptados para este novo paradigma informacional? Ainda, como sugere Ribeiro (2014) "É uma tecnologia? Uma ferramenta? Uma metodologia nova? Como o profissional da informação pode se inserir na discussão deste tema?".

Os dados existentes nos arquivos são uma importante fonte de análise, que dependendo do volume, pode necessitar de uma abordagem a partir do *Big Data*. Partimos de bancos de dados e simples tabelas do Excel com dados ordenados, avançando para inúmeros formatos. O DROID é uma ferramenta de software desenvolvida pelo The National Archives (www.nationalarchives.gov.uk) para realizar a identificação automatizada de formatos de arquivos de objetos digitais, identificando, atualmente, cerca de mil diferentes formatos de arquivos digitais, dentre eles extensões de arquivos de texto, imagem, áudio, vídeo, dados de GPS, informações na *Web*, dados de antenas e sensores, banco de dados relacionais, *flash*, SMS, mensagens instantâneas e tantos outros.

Para captura e uso de dados não se pode mais impor um formato de entrada, no sentido em que novas aplicações surgem a todo momento, assim como novos formatos de dados ganham vida. Sendo assim, variedade em *Big Data* refere-se a todos os dados estruturados e não estruturados que tem a possibilidade de serem gerados por seres humanos ou por máquinas. Os dados mais comumente adicionados são textos, *tweets*, fotos e vídeos estruturados. No entanto, dados não estruturados, como e-mails, correios de voz, texto manuscrito, gravações de áudio, etc., também são elementos importantes da variedade. O Google, por exemplo, usa *smartphones* como sensores para determinar as condições de tráfego em qualquer lugar do mundo. Nesta aplicação é provável que eles consigam ler a velocidade e a posição de milhões de carros para construir o padrão de tráfego, a fim de selecionar as melhores rotas. Esse tipo de dado não existia em escala coletiva há alguns anos.

Na busca por estas respostas nos debruçamos em pesquisar os autores, textos e referências que estivessem escrevendo sobre *Data Scientist*, nomenclatura dada ao profissional que atua com *Big Data*. DJ Patil é indiscutivelmente o mais conhecido cientista de dados do mundo. Ele foi cientista chefe de dados da Casa Branca; construiu a primeira equipe de ciência de dados no LinkedIn e, junto com Jeff Hammerbacher, é creditado pela Forbes por ter cunhado o termo "cientista de dados"[14]: "É um profissional de alto nível com o treinamento e a curiosidade para fazer descobertas no mundo do *Big Data*" (Davenport; Patil, 2012: 72).

Não obstante, os cientistas de dados fazem descobertas enquanto navegam em dados. A curiosidade, característica principal do cientista de dados, faz com que este profissional fique à vontade no mundo digital, sendo capazes de estruturar grandes quantidades de dados e possibilitar sua análise. Os cientistas de dados ajudam na tomada de decisão ao fazer análise contínua de dados, fornecendo subsídios para os tomadores de decisão, num cenário competitivo em que os desafios continuam mudando e os dados nunca param de fluir. Uma importante ferramenta de crescimento nas mãos de um profissional. Ao fazer descobertas, eles comunicam o que aprenderam e sugerem suas implicações para novas direções de negócios. Frequentemente, são criativos em exibir informações visualmente e fazer com que os padrões sejam claros e convincentes. Eles aconselham executivos e gerentes de produto sobre as implicações dos dados para produtos, processos e decisões. Davenport e Patil (2012: 73, tradução nossa) questionam "Que tipo de pessoa faz tudo isso? Que habilidades tornam um

14 https://www.forbes.com/pictures/efej45gkjj/2-jeff-hammerbacher-chief-scientist-cloudera-and-dj-patil-entrepreneur-in-residence-greylock-ventures/#634e9df32e53

cientista de dados bem sucedido? Pense nele ou nela como um híbrido de hacker, analista, comunicador e mentor. A combinação é extremamente poderosa e rara".

Song; Zhu (2017), no artigo intitulado *Big Data and Data Science: Opportunities and Challenges of iSchools*, discorrem sobre a necessidade de reavaliar os currículos das *Informations Schools*. Os autores defendem que os alunos das iSchool devem conhecer e aprender sobre as inúmeras tecnologias e ferramentas que contribuem para o desenvolvimento digital, tais como: Spark, NoSQL, NewSQL, virtualização de dados, análises de *Big Data*, *data lake*, *storage*, Internet das Coisas (IoT), inteligência artificial, robótica, realidade virtual, realidade aumentada, computação cognitiva, entre outras. Essas tecnologias estão cada vez mais presentes nos setores industriais e na sociedade como um todo. Os autores seguem dizendo que os alunos das escolas de ciência da informação, não precisam, necessariamente, aprender aspectos técnicos de todas essas tecnologias, mas eles precisam entender os conceitos, bem como seu papel enquanto profissional da informação quando na atuação com estas tecnologias, devem conhecer seus pontos fortes, aplicações e limitações tecnológicas.

O ambiente de *Big Data* pode ser especialmente desafiante para a Arquivologia, o que solicita uma reflexão sobre seus métodos e abordagens. Por um lado, os avanços tecnológicos têm impactado em toda a dinâmica de acesso e difusão de arquivos, de uma perspectiva interdisciplinar e tendo em conta seus elementos principais: os usuários, o conteúdo e a tecnologia (Rockembach, 2015). Além disto, a forma de análise e representação arquivística ganha novos contornos com as potencialidades oferecidas com o intercâmbio oferecido pela Ciência de dados:

> *Em uma era de dados digitais abundantes, os arquivistas estão cada vez mais pressionados para gerenciar arquivos sem essas técnicas [de análise de dados]. Além disso, essas técnicas oferecem muitas vantagens que as abordagens tradicionais de arquivamento não oferecem. Tomemos, por exemplo, o potencial das ciências de dados, como a visualização de informações, para transformar os processos de preservação arquivística ou a oportunidade de aplicar a análise visual para transformar representações arquivísticas. (Marciano et.al, 2018)*

Neste sentido, e corroborando com os autores, entendemos que o profissional da informação deve estar investido de conhecimento e, portanto, preparado para atuar com estas tecnologias, permitindo que eles se envolvam com a elaboração e uso de aplicativos e seus dados.

5. Considerações finais

Pretendemos com este artigo fazer uma análise sumária sobre o que se tem produzido cientificamente sobre *Big Data* e o cientista de dados, por entender que este nicho pode também ser assumido por Arquivistas, Cientistas da Informação e demais profissionais da Informação.

O volume, a velocidade e a variedade em que os dados estão sendo produzidos a todo momento são reflexos da explosão informacional vivida na atualidade, e disso não se tem dúvida. Porém, está cada vez mais latente a necessidade de encontrar soluções que auxiliarão na recuperação das informações. O *Big Data* é uma realidade e a cada novo ciclo vemos mais pesquisas sobre essa temática, criação de cursos especializados e ingresso da terminologia nos currículos acadêmicos.

Antes de discorrer sobre essa temática, se verificou a necessidade de compreender quem são e o que dizem os autores que discorrem sobre o tema. E, para isso, com a intenção de clarificar algumas questões, desenvolvemos uma estrutura conceitual sobre o *Big Data*, buscando entender quais são

seus principais conceitos e características, desde a contextualização a pontos importantes que constituem a área de investigação. A partir disso, entendemos que o uso dos grandes dados, associados à sua mineração, interpretação e uso, estava associado a um outro conceito: o de *datafication* ou dataficação, o qual procuramos definir e dimensionar. Verificamos que o poder não está nos dados armazenados, mas em sua mineração e seu uso. Os dados minerados são uma grande fonte de recursos utilizados com a intenção de monetizar a captação de todos os rastros que deixamos através do uso de sites de redes sociais e tantas outras plataformas online.

A análise dos artigos científicos se deu a partir da base Web of Science e os dados foram interpretados com auxílio do software BibExcel, que auxiliou na análise de dados bibliográficos. A partir da extração das informações sobre os artigos indexados, nos deparamos com as questões fundamentais levantadas pela área, que se relacionam principalmente com a disrupção que a análise destes grandes conjuntos de dados pode trazer para diversas áreas do conhecimento.

O trabalho trouxe contribuições significativas para compreender o estado atual de investigação, tanto no tocante ao *Big Data*, quando da necessidade de aprimorar a formação dos profissionais da informação para que comecem a atuar de forma mais relevante no mercado dos grandes dados. Importante, também, foi visualizar quais são os artigos mais importantes sobre o tema.

É inevitável dizer que o *Big Data* está cada vez mais presente na nossa Sociedade e que, a partir de sua existência em nosso cotidiano, temos a nossa frente um grande campo de estudo e atuação para os profissionais da informação.

Referências Bibliográficas

BOYD, D., & CRAWFORD, K. (2012) Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.

DAVENPORT, T. H., & PATIL, D. J. (2012) Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce. *Harvard Business Review*, 70.

DUMBILL, E. (2012) *What is big data? An introduction to the big data landscape*. Strata 2012: Making Data Work.

GANDOMI, A., & HAIDER, M. (2015) Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, *35*(2), 137-144.

GARTNER (2018) IT Glossary. Disponível em: http://www.gartner.com/it-glossary/big-data.

HAMPTON, S. E., STRASSER, C. A., TEWKSBURY, J. J., GRAM, W. K., BUDDEN, A. E., BATCHELLER, A. L., & PORTER, J. H. (2013) Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156-162.

JENKINS, H. (2009) *Cultura da convergência*. Aleph.

LYCETT, M. (2013) 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems*, *22*(4), 381-386.

MAI, J. E. (2016) Big data privacy: The datafication of personal information. *The Information Society*, *32*(3), 192-199.

MARCIANO, R., LEMIEUX, V., HEDGES, M., ESTEVA, M., UNDERWOOD, W., KURTZ, M., & CONRAD, M. (2018) Archival records and training in the age of big data. In *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education* (pp. 179-199). Emerald Publishing Limited.

MAYER-SCHÖNBERGER, V., & CUKIER, K. (2013) *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

MATTOSO, M. (2013) Scientific Workflows and Big Data. Palestra apresentada no *1º EMC Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro.

MCAFEE, A., BRYNJOLFSSON, E., DAVENPORT, T. H., PATIL, D. J., & BARTON, D. (2012) Big data: the management revolution. *Harvard Business Review*, *90*(10), 60-68.

NERURKAR, M., Wadephuland, C., & WIEGERLING, K. (2016) Ethics of Big Data: Introduction. *International Review of Information Ethics*, 24, 2-4.

RIBEIRO, C. J. S. (2014) Big Data: os novos desafios para o profissional da informação. *Informação & Tecnologia*, *1*(1), 96-105.

ROCKEMBACH, M. (2015) Difusão em arquivos: uma função arquivística, informacional e comunicacional. *Informação Arquivística*, *4*(1), 98-118

SONG, I. Y., & ZHU, Y. (2017) Big data and data science: opportunities and challenges of iSchools. *Journal of Data and Information Science*, *2*(3), 1-18.

SWAN, M. (2013) The quantified self: Fundamental disruption in big data science and biological discovery. *Big data*, *1*(2), 85-99.

TECHAMERICA FOUNDATION'S FEDERAL BIG DATA COMMISSION. (2012) *A Practical Guide to Transforming the Business of Government*. TechAmerica Foundation's Federal Big Data Commission.

VAN DIJCK, J. (2014) Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197-208.

WEERKAMP, W., & DE RIJKE, M. (2012, August) Activity prediction: A twitter-based exploration. In *SIGIR workshop on time-aware information access*.

ZIKOPOULOS, P., & EATON, C. (2011) *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.