XIX encontro nacional de pesquisa em ciência da informação
ENANCIB
// SUJEITO INFORMACIONAL E AS PERSPECTIVAS ATUAIS EM CIÊNCIA DA INFORMAÇÃO. //
22-26 OUTUBRO 2018
LONDRINA/PR
Universidade Estadual de Londrina

XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2018

GT-8 – Informação e Tecnologia

Evolução do tratamento e coleta de dados na ciência: E-science, BigData e Big Science

Valmir Batista Prestes de Souza (Doutorando Universidade de São Paulo - USP)

Marcos Luiz Mucheroni (Universidade de São Paulo – USP)

Cristiane Aparecida de Massena

Evolution of treatment and data collection in science: E-science, BigData e Big Science

Modalidade da Apresentação: Pôster

Resumo: Este trabalho é derivado de tese de doutorado onde são problematizadas as formas de coleta e tratamento de trabalhos acadêmicos de artigos, teses e dissertações nos campos da E-science, Big Data e Big Science relacionando sua evolução e contexto de publicação. Discute-se a importância e a fundamentação destas novas formas de tratamento e coleta de trabalhos científicos e sua importância na área da Ciência da Informação. A partir da coleta são analisados os resultados e impactos destas novas áreas de estudo e a complexidade do volume de trabalhos que foram coletados. Verifica-se a consolidação do tratamento de Big Data para trabalhos científicos, uma evolução nos trabalhos de e-Science e uma evolução ainda tímida de Big Science, porém que não deixa de ter relevância para a CI.

Palavras-Chave: Big Data, e-Science, Big Science, Tecnologia da Informação

Abstract: This work derives from a doctoral thesis where the ways of collecting and treating academic articles, theses and dissertations in the fields of E-science, Big Data and Big Science are related to their evolution and context of publication. It discusses the importance and the basis of these new forms of treatment and collection of scientific works and their importance in the area of Information Science. From the collection, the results and impacts of these new study areas and the complexity of the volume of work that was collected are analyzed. The consolidation of the treatment of Big Data for scientific works, an evolution in e-Science works, is still a timid evolution of Big Science, but is not without relevance for Information Science.

Keywords: Big Data, e-Science, Big Science, Information Technology.

1 INTRODUÇÃO

Durante a segunda metade do século XX, o investimento maciço em pesquisas científicas criou um fenômeno utópico na ciência, a chamada *Big Science*, que para teve sua origem com Rutherford, com grandes investimentos e pesquisadores, laboratórios nacionais e, por vezes, ligados a questões militares, como o Projeto Manhattan, avaliado em 30 bilhões de dólares, que coordenou a construção das bombas atômicas lançadas sobre o Japão em agosto de 1945. (VIEIRA, 2015)

A *Big Science* trouxe muitas tecnologias oriundas da química e física aplicadas em escala comercial, e nas últimas décadas surgiram aplicações como o microscópio eletrônico de tunelamento com varredura, laser e espectroscopia de nêutrons, os detectores de partículas elementares, e os circuitos integrados os quais renderam 8 prêmios Nobel de Física a seus inventores. (AGUIAR, 2018).

Atualmente, graças a investimentos estatais e privados, existem os grandes aceleradores que trazem uma nova vertente na ciência, intitulado *New Big Science*. Tem-se como exemplo o bóson de Higgs e o Grande Colisor de Hádrons (LHC), que envolve quase 10 mil pessoas de mais de cem países, e investimento de 5 bilhões de euros (BARROSO, 2008).

A maior diferença entre o *e-science* e o *Big Science* é que as iniciativas de *e-science* costumam privilegiar o desenvolvimento de uma ciência aberta (*open science*), ligada principalmente à disponibilização e manutenção de bases de dados abertos, de acesso público, que subsidiem o trabalho de pesquisa tanto no âmbito individual como no colaborativo (APPEL, 2014).

2 *E-Science e Big Data*

A crescente necessidade de tratar os dados se tornou tão especializado que é uma ciência própria com seus saberes. É multidisciplinar e demanda metodologia específica para coletar, analisar e visualizar dados, para extrair *insights* e informações dos dados para tomar decisões e fazer previsões (NIELSEN; BULINGAME, 2012).

A *Data Science* permite às empresas a capacidade de transformar seus ativos de dados em insumos para resolver problemas complexos e criar estratégias mais inteligente. A chave é a adição de valor com o aprendizado sobre os dados (MATOS, 2017).

O crescente volume de dados e o conjunto de soluções tecnológicas para trata-los gerou o *Big Data*. Para Mayer-Schonbeerger e Cukier (2013), o *Big Data* representa "uma nova fonte de valor econômico e informação".

Para Gartner (2018), o termo *Big Data*, compreende um grande volume de "informações e ativos de informações de alta velocidade e ainda a variedade que exigem formas inovadoras e econômicas de processamento, permitindo uma visão aprimorada, tomada de decisão, e automação de processos''. Canary (2013) reuniu algumas definições segundo diversos autores, como pode-se observar no quadro a seguir :

Quadro 1 – referencial do *Big Data*.

MANYKA, J, et. al.(2011) (McKinsey Global Institute)	*Big Data* refere-se a conjuntos de dados cujo tamanho é além da capacidade de ferramentas de *software* de banco de dados típicos para capturar, armazenar, gerenciar e analisar
MCAFEE, A; et. al. (2012) (Harvard Business Review)	*Big Data* como uma forma essencial para melhorar a eficiência e a eficácia das organizações de vendas e *marketing*. Ao colocar *Big Data* no coração de vendas e *marketing*, os *insights* podem ser aproveitados para melhorar a tomada de decisão e inovar no modelo de vendas da empresa, o que pode envolver as utilização de dados para orientar ações em tempo real
DEMIRKAN, et. al. (2012) (Decision Support Systems)	Há o desafio de gerenciar grandes quantidades de dados (Big Data), que está ficando cada vez maior por causa do armazenamento mais barato e evolução dos dados digitais e dispositivos de coleta de informações, como telefones celulares, laptops, e sensores
PHELAN, Mike (2012) (Forbes)	O fenômeno surgiu nos últimos anos devido à enorme quantidade de dados da máquina que está sendo gerao hoje – [..] – juntamente com as informações adicionais obtidas por análise de todas essas informações, que por si só cria outro conjunto de dados enorme
Gartner Group (2012)	Big Data, em geral, é definido como ativos de alto volume, velocidade e variedade de informação que exigem custo-benefício, de formas inovadoras de processamento de informações para maior visibilidade e tomada de decisão
Internacional Data Corporation	As tecnologias de Big Data descrevem uma nova geração de tecnologias e arquiteturas projetadas para extrair economicamente o valor de volumes muito grandes e de uma grande variedade de dados, permitindo alta velocidade de captura, descoberta, e/ou análise

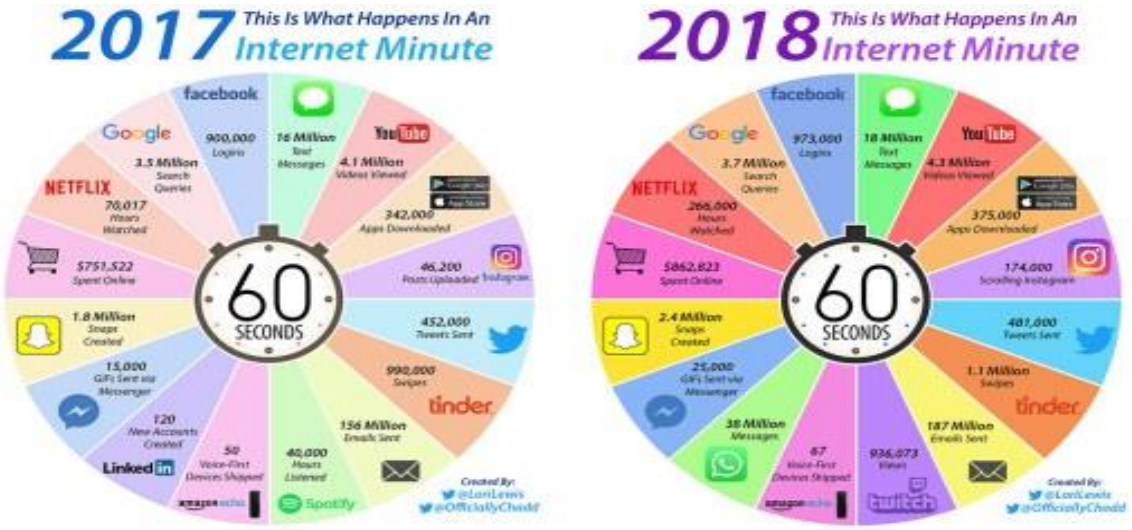
Fonte: Adaptado pelos autores de Canary (2013)

Segundo Marr (2015) o *Big Data* afetará todos os negócios e vidas e até 2020, cerca de 1,7 megabytes de novas informações serão criadas por segundo para cada pessoa no planeta. Existem dimensões que definem o big Data em 10vs:

Volume: a grande quantia de dados que devido ao seu volume se apoia em computação distribuída e *softwares* apropriados a este tipo de cenário (TAURION, 2012).

O infográfico a seguir mostra os dados dos principais *sites* e acontece na *internet* em um minuto e exemplifica em números o *Big Data* no ano de 2017 e seu crescimento em relação a 2018:

Figura 1 – Infográfico – a evolução da infosfera e do *Big Science*.



Velocidade para Taurion (2012) está relacionada à maneira como os dados são gerados rapidamente. Para Gartner (2018), velocidade significa tanto o quão rápido os dados estão sendo produzidos e tratados para atender a demanda.

Variedade: A variedade do *Big Data* é formada por uma crescente pluralidade de fontes e diversidade de dados que devem ser tratadas para processamento e análise. (TAURION, 2012).

Veracidade: há ferramentas que analisam a veracidade dos dados cuja origem possa ser pouco confiável com o propósito validar sua origem evitando fontes de *fake news* (notícias falsas).

Valor: possivemente é a principal variável a ser avaliada. Se não é possível transformar estes dados em valor, os mesmos não terão utilidade.

Variabilidade é o número de inconsistências, como anomalias e *outliers* antes que ocorra qualquer análise dos dados.

Validade: refere-se à precisão e correção dos dados para o uso pretendido.

Vulnerabilidade: A violação de dados neste universo também atinge grandes proporções.

Volatilidade: quanto tempo se deve preservar o dado é útil antes que seja irrelevante, pois o custo de preservar e disponibilizar os dados e sua recuperação devem ser considerados.

Visualização: existe o desafio para visualizar tantos dados pelas limitações da tecnologia (FIRICAN, 2017).

3 Método

Trata de estudo quantitativo das teses, dissertações e artigos científicos que envolvam *e-Science* e *Big Science*, onde o tratamento e a coleta de informações envolvem um número maior de dados e já ultrapassaram em muito os limites da era da explosão informacional, e também as formas de recuperação e tratamento de dados necessitam de atualização. Os dados foram coletados a partir dos repositórios oasisbr[1] (IBICT) e catálogo de teses[2] (CAPES) que compreende a produção científica de autores vinculados a universidades e institutos de pesquisas brasileiros.

Os parâmetros utilizados na busca foram "*e-Science*" ou *escience* e "*Big Science*", sendo coletados os dados individualmente para cada termo da pesquisa, visto que a junção de mais de um termo, no catálogo de teses (CAPES) gerava resultados divergentes, superiores à soma das buscas individuais.

4 Resultados

Atualmente, a *Data Science* é vista como a teoria e a prática de extrair conhecimento de dados e se preocupa com a criação de produtos e/ou serviços a partir de dados e resolver problemas reais de negócios, com o uso de método científico e técnicas avançadas de análise de dados, *machine learning* e inteligência artificial.

Nessa perspectiva, a presente trabalho identificou as pesquisas científicas que abordam os temas *e-Science, Big Science* em artigos, dissertações e teses. Para tanto, foi realizada coleta nos repositórios Oasis (IBICT) e catálogo de teses (CAPES), visto que algumas pesquisas constam em apenas um dos repositórios, conforme quadro a seguir:

Quadro 2 – Produção Científica

Produção	Catálogo de Tese – CAPES		Oasisbr	
	e-Science	*Big Science*	*e-Science*	*Big Science*
Mestrado	35	3	5	3
Doutorado	29	1	4	4
Artigo	0	0	43	30

Fonte: elaborado pelos autores

[1] Portal Brasileiro de publicações científicas em acesso aberto http://oasisbr.ibict.br/vufind/

[2] Catálogo de Teses e Dissertações CAPES. Disponível http://catalogodeteses.capes.gov.br/catalogo-teses/

Diante de temas que vem sendo discutidos e pesquisados, a tabela 2 extratifica as pesquisas científicas no panorama nacional que possuem relação direta e indireta em *Big Science* e *e-Science.*

A produção científica direta corresponde às pesquisas nos temas *Big Science* e/ou *e-Science* em artigos, dissertações e teses, conforme tabela 3; já as produções indiretas correspondem as pesquisas de outras temáticas mas estão vinculadas a grupos de pesquisas que trabalham com os referidos temas.

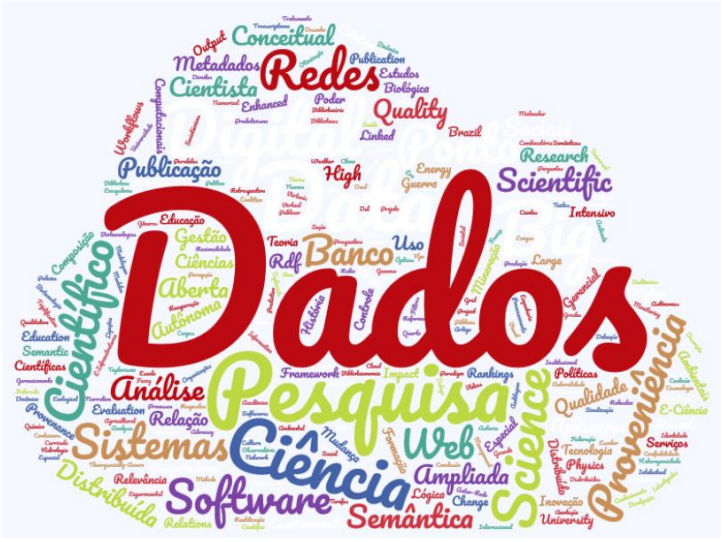
Quadro 3 – Produção Científica Direta

Produção	Catalogo de Tese – CAPES		Oasisbr	
	e-Science	Big Science	e-Science	Big Science
Mestrado	22	2	4	1
Doutorado	17	0	3	0
Artigo	0	0	13	18

Fonte: elaborado pelos autores

Para identificar os artigos, dissertações e teses nos repositórios utilizou-se de busca no título, resumos e palavras-chaves, e para ilustrar os termos encontrados nas palavras-chaves, a seguir a nuvem de palavras identificadas na pesquisa.

Figura 2 – Nuvem de palavras-chaves



Fonte: elaborado pelos autores

CONSIDERAÇÕES FINAIS

A explosão informacional impulsiona a utilização de recursos computacionais que contribuem no armazenamento, manuseio, tratamento e comunicação científica das pesquisas nacionais. Dessa forma, os conceitos apresentados de *Big Data, Big Science* e *E-Science* tornam-se essenciais e corroboram com a difusão, produção e disseminação científica.

Foi possível observar que as pesquisas sobre *Big Science* e *e-Science* também abordam *Big Data*, compartilhamento de dados, pesquisa colaborativa, *web* semântica e Banco de Dados. Mas a pesquisa identificou que tanto o repositório mantido pelo IBICT quanto o mantido pela CAPES, que deveriam conter as pesquisas nacionais de mestrado e doutorado, deixam a desejar, pois não contêm todas as pesquisas elaboradas nos programas de Pós-Graduação, conforme os resultados apresentados na tabela 2.

Por fim, para uma pesquisa futura, pretende-se buscar no Currículo *Lattes* os pesquisadores que desenvolvem pesquisas sobre *Big Data, Big Science* e *e-Science*, mapeando as instituições, programas de Pós-Graduação e a rede de colaboração de pesquisadores nacionais e confrontar com os resultados aqui obtidos.

REFERÊNCIAS

AGUIAR, Ricardo. **A Nova Big Science.** Disponível em: <https://www.sprace.org.br/divulgacao/noticias/a-nova-big-science>. Acesso em: 25 jul. 2018.

APPEL, André L. **A e-Science e as atuais práticas de pesquisa científica.** Dissertação (Mestrado em Ciência da Informação). Rio de Janeiro: Instituto Brasileiro de Informação em Ciência e Tecnologia, 2014.

BARROSO, Sérgio. O grande da Big Science. **Rev**. **Princípios**, out-nov, 1998. Disponível em: <http://revistaprincipios.com.br/artigos/98/cat/675/o-grande-da-&quotbig-science%22-.html>. Acesso em 25 jul. 2018

BUFREM, Leilah S; et al. Produção Internacional Sobre Ciência Orientada a Dados: análise dos termos Data Science e E-Science na Scopus e na Web of Science. **Informação & Informação**, Londrina, v. 21, n. 2, p. 40-67, dez. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/26543>. Acesso em: 25 jul. 2018.

CANARY, Vivian P. **A tomada de decisão no contexto do big data:** estudo de caso único. Disponível em: <https://lume.ufrgs.br/handle/10183/87757>. Acesso em 25 jul. 2018

FIRICAN, G. The 10 Vs of Big Data. **TDWI**, 2017. Disponível em: <https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx>. Acesso em: 25 jul. 2018.

GARTNER. **Glossário**. Disponível em: <https://www.gartner.com/it-glossary/big-data>. Acesso em 22 jul 2018

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 1991.

KNIGHT, M. What is data Science?. Dataversity, 13 nov. 2017. Disponível em: <http://www.dataversity.net/what-is-data-science/>. Acesso em 15 jul 2018.

MARR, Bernard. Big Data: 20 Mind-Boggling Facts Everyone Must Read. FORBES, Sep 30, 2015. Disponível em: <https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#6b8ba1b217b1>. Acesso em 25 jul. 2018

MATOS, David. Ciência de Dados e soluções. Disponível em <http://www.cienciaedados.com/ciencia-de-dados-e-solucoes/>. Acesso em: 25 jul. 2018.

MAYER- SCHONBERGER, V; CUKIER, K. *Big data:* como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Elsevier, 2013.

NIELSEN, Lars; BURLINGAME, N. **A Simple Introduction to data Science.** Wickford: New Street Communications, 2012.

NUNES, Roberto C; DANTAS, J.M.; ANDRADE, J.E. Big Science e a formação do modelo de partículas elementares. **Scientia plena**, v. 13, n. 1, 2017. Disponível em: <https://www.scientiaplena.org.br/sp/article/view/3519>. Acesso em: 18 Jul. 2018.

TAURION, C. **Você realmente sabe o que é o big data?** Blog da IBM, 30 abril 2012. Disponível em: <https://www.ibm.com/developerworks/>. Acesso em: 22 jul. 2018.

VIEIRA, C.L. (org). **História da física**: artigos, ensaios e resenhas. Rio de Janeiro: Instituto Ciência Hoje, 2015 Disponível em: <http://www.nossaciencia.com.br/files/livro-historia-da-fisica.pdf>. Acesso em 25 jul. 2018.