DOI: 10.5433/1981-8920.2016v21n2p143

CURADORIA DIGITAL: PROPOSTA DE UM MODELO PARA CURADORIA DIGITAL EM AMBIENTES BIG DATA BASEADO NUMA ABORDAGEM SEMI-AUTOMÁTICA PARA A SELEÇÃO DE OBJETOS DIGITAIS

CURACIÓN DE CONTENIDOS DIGITALES: PROPUESTA DE UN MODELO PARA CURACIÓN DE CONTENIDOS DIGITALES EN AMBIENTES BIG DATA BASADO EN UN ENFOQUE SEMIAUTOMÁTICO DE SELECCIÓN DE OBJETOS DIGITALES

Moisés Lima Dutra[*]
Douglas Dyllon Jeronimo de Macedo[**]

RESUMO
Introdução: Expõe um novo olhar para Curadorias Digitais a partir de uma perspectiva Big Data.
Objetivo: Propõe técnicas de seleção e avaliação de objetos digitais para curadorias digitais que levem em conta o volume, a velocidade, a variedade, a veracidade e o valor dos dados coletados em múltiplos domínios do conhecimento.
Metodologia: Trata-se de uma pesquisa exploratória de natureza aplicada e de abordagem qualitativa. A aplicação de heurísticas permite que este processo seja feito tanto por curadores humanos quanto por agentes de software.
Resultados: Apresenta um modelo para busca, tratamento, avaliação e seleção de objetos digitais a serem tratados em curadorias digitais.
Conclusões: É possível utilizar ambientes Big Data como fonte de recursos informacionais para curadorias digitais. Técnicas e ferramentas Big Data podem auxiliar no processo de busca e seleção de recursos informacionais para serem tratados em Curadorias Digitais.

Palavras-chave: Curadoria Digital, Big Data, Objetos Digitais.

* Doutor em Ciência da Computação (Doctorat En Informatique) pela Université Claude Bernard Lyon 1 (UCBL). Professor Adjunto do Departamento de Ciência da Informação da Universidade Federal de Santa Catarina (UFSC).
**Doutor em Engenharia e Gestão do Conhecimento pela Universidade Federal de Santa Catarina (UFSC). Professor Adjunto do Departamento de Ciência da Informação da UFSC.

1 INTRODUÇÃO

Os impactos resultantes de cenários proporcionados por tecnologias que alteram substancialmente o nosso modo de viver não cessam de servir de objeto para novas pesquisas científicas. É necessário se encontrar respostas para questões do tipo: “Quais são as implicações a longo prazo do crescimento da geração ininterrupta e exponencial de novos dados?”, “Onde estes dados serão armazenados?”, e “Como eles serão processados, de maneira a poderem ser extraídos e recuperados?”. Estamos tão imersos neste mundo tecnológico regido por *gadgets* – que geram uma quantidade extraordinária de dados por minuto –, que, muitas vezes, não nos damos conta do esforço necessário para tornar esses processos e atividades operacionais. A grande quantidade de dados e informações presentes atualmente em bases de dados, geradas nos mais diversos domínios de aplicação, torna bastante complexa a tarefa de gerenciar todo este material. Além disso, há também a questão da manutenção destes dados por meio da aplicação de técnicas de preservação digital.

Conforme afirma Beagrie (2006), os desafios de instituições como bibliotecas e arquivos – que precisam pensar em séculos à frente –, na verdade acabam por se manifestar, num mundo digital, em uma década ou menos. Questões similares acabam por afetar igualmente as organizações de maneira geral. Novas regulações e marcos legais, prestações de contas em todos os setores, como por exemplo, em instituições bancárias, farmacêuticas, médicas e aeroespaciais, significam que estas organizações precisam guardar informações digitais que sejam acessíveis por, pelo menos, uma década ou mais. Para outros tipos de organizações, tais como as de entretenimento e mídia, seus ativos hoje em dia já são majoritariamente digitais. Num mundo influenciado cada vez mais por mudanças tecnológicas, percebe-se claramente o desafio existente na manutenção da perenidade de dados e informações nas organizações.

No entanto, os desafios digitais se aplicam também a pessoas, não apenas a organizações. As ferramentas colaborativas da Web 2.0 que, durante

a última década e meia têm permitido que o usuário comum da Internet produza, armazene e publique informações num ritmo nunca visto anteriormente, são mais um dos fatores que capitaneiam este processo. O custo relativamente baixo dos novos dispositivos móveis, a maior largura de banda disponibilizada pelos provedores de acesso e o virtual "custo zero" do armazenamento nas nuvens também possibilitam que cada vez mais informações digitais produzidas por indivíduos – tais como fotos, e-mails, documentos texto, planilhas, imagens, gráficos, vídeos, áudios, *tweets*, entre outros –, possam estar publicadas e à disposição de todos.

Em contextos de Curadorias Digitais (CD), que lidam com a preservação e com o gerenciamento do ciclo de vida de objetos digitais, este problema se torna ainda maior, pois que a ação de selecionar os objetos na CD é uma tarefa primordial deste processo. Além do mais, é necessário considerarmos que a existência dos múltiplos domínios de conhecimento, que servem de fontes de informação para as CD, implica em diferentes tipos de dados e formatos de representação destes e, consequentemente, em diferentes exigências. Percebe-se, portanto, a necessidade de se desenvolver novas técnicas para automatizar o processo de seleção de novos objetos digitais para os repositórios das CD. Neste processo estão inclusos a geração de metadados representativos das características do ciclo de vida dos objetos digitais, o desenvolvimento de novos de agentes de software e o estabelecimento de serviços colaborativos para o compartilhamento de conteúdo multipropósito.

Para Sayão e Sales (2013), tópicos como modelos e técnicas para processamento inteligente e de descoberta de dados por meio de taxonomias e ontologias, integração com os padrões da Web Semântica e do *Linked Data* também estão na ordem do dia no que diz respeito ao desenvolvimento das CD.

Há ainda questões relacionadas à segurança da informação, sobretudo no que tange aos quatro pilares clássicos de autenticidade, integridade, confidencialidade e disponibilidade, que têm sua complexidade agravada neste cenário.

Beagrie (2006, p.12, tradução livre), questiona:

> Como garantir a segurança física de um material por décadas a fio? Como proteger a privacidade? Como organizar e extrair conhecimento útil desta rica biblioteca informacional e usá-lo efetivamente? Com relação ao material que se pretende compartilhar, como efetivamente controlar o acesso a este por diferentes grupos de usuários?

O paradigma Big Data surgiu com o intuito de tratar questões contemporâneas complexas, originadas no aumento eminente dos volumes de dados, que são compostos por uma quantidade também crescente de tipos e formatos, que necessitam de velocidade ótima durante a sua recuperação. Big Data então consiste primariamente de conjuntos de dados, caracterizados pelo volume, variedade e velocidade, que necessitam de arquiteturas escaláveis para que os processos de armazenamento, manipulação e análise sejam eficientes (NIST, 2016). Este conceito foi expandido nos últimos anos para tratar mais dois aspectos primordiais, sendo estes a veracidade dos dados e o valor que podemos extrair destes dados, para a tomada de decisão.

Este trabalho propõe um modelo de uma arquitetura conceitual multidomínio, que tem como função interagir com múltiplos repositórios e bases de dados, e, a partir de heurísticas pré-definidas, selecionar objetos digitais que possuam valor para as Curadorias Digitais, facilitando assim, o processo de busca e seleção em grandes massas de dados. Ao final, nossa proposta é desenvolver um modelo que possa melhorar e otimizar o processo de seleção de objetos digitais em ambientes Big Data, de maneira que estes possam ser tratados por CDs. Como complemento, também será apresentado um modelo de operação que tem por objetivo facilitar o entendimento do funcionamento do modelo conceitual.

Duas perguntas de pesquisa guiam este trabalho. A primeira: "**É possível utilizar ambientes Big Data como fonte de recursos informacionais para Curadorias Digitais?**". A segunda: "**Como as técnicas e ferramentas Big Data podem auxiliar no processo de busca e seleção de recursos informacionais para serem tratados por Curadorias Digitais?**". Estas perguntas de pesquisa estão atreladas à problemática da geração

exponencial de dados dos últimos anos e ao desafio dos processos de busca e seleção de recursos informacionais/objetos digitais relevantes para serem preservados.

Embora na sua essência, a ideia de Curadoria Digital remeta a um processo de seleção e manutenção de objetos digitais, neste trabalho, para fins didáticos, iremos considerar também como parte integrante da CD os repositórios temáticos resultantes deste processo. A importância desta decisão se deve à necessidade de se representar o processo de seleção semiautomático de objetos digitais em diferentes domínios de conhecimento. Não é objeto deste trabalho, no entanto, procurar redefinir os conceitos associados à ideia de Curadoria Digital.

Este trabalho está estruturado em quatro seções, sendo esta a primeira delas. Na seção 2 é apresentado o referencial teórico do artigo, trazendo à luz os temas Curadoria Digital, com seu histórico, tópicos sobre preservação digital e possíveis área de aplicação. Ainda na seção 2, o tema Big Data é apresentado e detalhado quanto ao Volume, Velocidade, Variedade, Veracidade e Valor. Na seção 3, a proposta de modelo para Curadoria Digital em ambientes Big Data é apresentada. Por fim, na seção 4, as considerações finais do trabalho são delineadas.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão tratados os fundamentos teóricos do trabalho, visando delimitar os conceitos que embasam esta proposta. Aqui, serão tratados os temas de Curadoria Digital e Big Data.

2.1 Curadoria Digital

A ideia de “curadoria” nos remete quase que imediatamente ao termo utilizado de maneira mais tradicional por museus e bibliotecas, especialmente em relação a coleções de artefatos físicos.

Em sua etimologia, o termo curadoria está vinculado ao ato de curar, zelar, vigiar por algo: um conceito originalmente relacionado aos campos do

Direito e das ordens monásticas. Com a evolução social, o termo passa a relacionar-se com o campo das artes, dos museus, das bibliotecas e de seus respectivos acervos (CORRÊA; BERTOCCHI, 2012).

> Por curadoria podemos compreender o conjunto de ações que garantem que um conjunto de dados é genuíno, permitindo o seu uso por outros que não os seus produtores. A curadoria pode envolver ações de descrição dos dados, de ligação destes a outros que os tornem inteligíveis, de registo dos usos que tenham e dos resultados a que tenham dado origem (FERREIRA *et al*., 2012, p.26).

Abbot (2008) define Curadoria Digital como o conjunto das todas as atividades existentes no gerenciamento de dados, desde o planejamento da sua criação, passando pela digitalização (em caso de materiais analógicos) ou criação (para os materiais já gerados em meio eletrônico), procurando assegurar a disponibilidade e adequação para a recuperação e reuso futuro destes dados. A autora também afirma que a CD inclui o gerenciamento de uma vasta quantidade de *datasets* de uso diário, visando garantir que os mesmos estejam visíveis e acessíveis, e que assim continuem. Para ela, a ideia de CD é aplicável a uma vasta gama de atividades profissionais, do início ao fim do ciclo de vida dos objetos digitais, sendo passível de ser trabalhada por: digitalizadores, criadores de metadados, financiadores de projetos, legisladores, gestores de repositórios digitais, entre outros. Yakel (2007) declara que diversos conceitos estão relacionados às CD, entre os quais podemos destacar:

a) Gerenciamento do ciclo de vida dos materiais desde a sua criação;
b) Interação ativa ao longo do tempo entre os criadores de materiais e os potenciais curadores digitais;
c) Avaliação e seleção de materiais (*appraisal*);
d) Desenvolvimento e disponibilização de acesso a estes materiais;
e) Garantia de preservação (usabilidade e acessibilidade) dos objetos digitais.

Na última década, o termo Curadoria Digital capturou o imaginário de uma crescente comunidade internacional de profissionais e pesquisadores, advindos das mais diversas áreas (DALLAS, 2015). Para Sayão e Sales (2012, p.185),

> [...] a curadoria digital emerge como uma nova área de práticas e de pesquisa de espectro amplo que dialoga com várias disciplinas e muitos gêneros de profissionais. Ela une as tecnologias e boas práticas do arquivamento e da preservação digital e dos repositórios digitais confiáveis com a gestão dos dados científicos, criando uma nova área de pesquisa cujos desdobramentos, de amplo espectro, ainda são imprevisíveis. Isto porque, como se trata de uma área que só recentemente despontou como crítica para a pesquisa, ainda restam muitas lacunas práticas e teóricas a serem equacionadas, orientadas, preferencialmente, por uma abordagem multidisciplinar.

A grosso modo, Curadoria Digital "diz respeito à **manutenção** e à **criação de valor** em conjuntos confiáveis de informações digitalizadas para **uso corrente** e **futuro**" (GIARETTA, 2005 *apud* BEAGRIE, 2006, p.6, tradução livre, grifos do original). Para Beagrie (2006), inclusive, o termo CD implica não apenas na preservação e manutenção de uma coleção ou base de dados, mas em algum grau de valor agregado e conhecimento.

2.1.1 Histórico do Termo

"Curadores digitais são o novo tipo de profissional existente nas redondezas" (YAKEL, 2007, p.335, tradução livre). Apesar dos termos "digital", "curador" e "curadoria" já fazerem parte do vocabulário dos profissionais da informação há décadas, só recentemente é que surge o termo Curadoria Digital.

Na metade dos anos 1990, as ações de preservação digital no Reino Unido se concentravam na sobrevivência do material digital, sobretudo estimuladas pelo relatório norte-americano sobre preservação digital, *Task Force on Archiving of Digital Information*, de 1996 (HIGGINS, 2011). Com o tempo, novos desenvolvimentos técnicos e uma compreensão mais madura da atividade organizacional e do fluxo de trabalho deslocaram a ênfase dessas

ações para a garantia do acesso, uso e reuso de materiais digitais ao longo do ciclo de vida destes. Para Kim (2014, p.1, tradução livre), esta comunidade "alargou suas fronteiras de uma preservação passiva para uma curadoria ativa".

O termo Curadoria Digital propriamente dito foi utilizado pela primeira vez em um seminário sobre e-Science, arquivos e bibliotecas digitais (*Digital Curation: digital archives, libraries and e-science seminar*), realizado em Londres, em outubro de 2001 (BEAGRIE, 2006).

Yakel (2007) considera 4 trabalhos como sendo chave para o desenvolvimento da CD em suas origens:

> 1) O relatório de Dan Atkins do National Science Foundation (NSF) de 2003, chamado "Revolucionando a ciência e a engenharia por meio da ciberinfraestrutura: relatório da NSF para o painel consultivo sobre ciberinfraestrutura";
> 2) O relatório do American Council on Learned Societies (ACLS) de 2006, intitulado "Nossa comunidade cultural: relatório final da Comissão do ACLS para ciberinfraestrutura em Humanidades e Ciências Sociais";
> 3) O relatório do Conselho de Ciberinfraestrutura do National Science Foundation (NSF), de 2007, intitulado "Visão ciberinfraestrutural do NSF para o Século 21";
> 4) O relatório de Liz Lyon de 2007, intitulado "Lidando com dados: papéis, direitos, responsabilidades e relações".

Para Higgins (2011), todos estes fatores contribuíram para que a CD se manifestasse como uma nova disciplina, baseada em atividades do Digital Curation Centre[1] (DCC) e em diversos projetos desenvolvidos no escopo do programa de financiamento de pesquisa EU 6th Framework, da Comissão Europeia, que se estendeu de 2002 a 2006.

Em 2005, surge a primeira conferência de grande porte sobre o tema, a International Digital Curation Conference[2], que aconteceu em Bath, no Reino Unido, e que se tornou um evento anual desde então (KIM, 2014). Para efeitos comparativos, a primeira conferência internacional sobre preservação digital, a

[1] http://www.dcc.ac.uk/
[2] http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc

International Conference on Digital Preservation[3] havia ocorrido em 2004, em Pequim, na China.

Um workshop chamado de *Digital Curation & Trusted Repositories, Seeking Success* foi realizado na Joint Conference on Digital Library[4] de 2006 (DALLAS, 2015). O autor diz ainda que nas reuniões que ocorreram durante este evento, o termo CD foi apresentado como uma nova área de especialização em departamentos de Ciência da Informação de diversas universidades, entre as quais se destacam: a Universidade da Carolina do Norte (UNC), a Universidade de Illinois (U of I), a Universidade de Siracusa (SU) e a Universidade do Norte do Texas (UNT).

Neste mesmo ano de 2006, surge também o periódico International Journal of Digital Curation[5] (IJDC). O IJDC é um periódico eletrônico de acesso aberto publicado em duas edições por ano que, segundo a definição encontrada no seu website, é inteiramente devotado a *papers*, artigos e novidades relacionadas a curadoria digital, objetos digitais e questões relacionadas.

De maneira geral, pode-se dizer que o desenvolvimento da CD tem se dado nos últimos anos sobretudo no Reino Unido e nos Estados Unidos. Com o apoio financeiro de políticas implementadas por agências como o Joint Information Systems Committee, do Reino Unido, de programas de fomento como o 6th e o 7th Frameworks da Comissão Europeia e do Institute of Museum and Library Services (IMLS), dos Estados Unidos, a CD forma um extenso ecossistema de bases institucionais, inciativas e projetos de pesquisa, programas e currículos de especialização profissionais e infraestrutura de serviços e ferramentas digitais (KIM, 2014). O termo, no entanto, vem ganhando importância nos últimos anos, sobretudo por se relacionar intimamente com a questão da preservação digital.

[3] https://ipres-conference.org/
[4] http://www.jcdl.org/
[5] http://www.ijdc.net/

2.1.2 Preservação Digital

Para Yakel (2007), o termo Curadoria Digital acabou por se tornar um guarda-chuva para vários outros conceitos associados, tais como preservação digital, curadoria de dados, gestão de recursos eletrônicos, gestão de ativos, entre outros. Kim (2014) confirma esta afirmação na pesquisa que empreendeu no conteúdo de artigos do IJDC entre 2006 e 2013. Para ele, o campo da Curadoria Digital é um domínio híbrido do conhecimento, que engloba diversas áreas relacionadas ao gerenciamento e à preservação de objetos digitais ao longo do tempo. Para Ferreira *et al*., (2012), a curadoria envolve também ações de preservação, em que a representação dos dados e os seus metadados tenham de ser modificados.

Os dados armazenados em formato eletrônico têm necessidade de um arcabouço tecnológico (*software* e/ou *hardware*) que faça a transição dos *bits* para formatos humanamente compreensíveis e operacionais. A grande questão é que o contexto tecnológico que produz esses arcabouços é extremamente volátil e suscetível à obsolescência. Desta forma, é necessário se investir em técnicas e ferramentas que adaptem os arcabouços tecnológicos mais recentes a contextos já obsoletos.

Para Ferreira *et al*. (2012, p.9), "a preservação digital preocupa-se com a capacidade de manter a informação digital acessível, interpretável e autêntica, mesmo na presença de uma plataforma tecnológica diferente daquela inicialmente utilizada no momento da sua criação".

Nos últimos anos, pôde-se perceber avanços em relação à compreensão da ideia de preservação digital. Em 2004, Arellano afirmava que:

> Com o aumento da produção de informação em formato digital, tem sido questionada cada vez mais a importância de se ter garantida a sua disponibilização e preservação por longos períodos de tempo. Essa preocupação envolve tanto os produtores dos dados quanto os órgãos detentores dessa informação. No início, as práticas relacionadas com a preservação digital estavam baseadas na idéia de garantir a longevidade dos arquivos, mas essa preocupação agora está centralizada na ausência de conhecimento sobre as estratégias de preservação digital e o que isso poderá significar na

> necessidade de garantir a longevidade dos arquivos digitais (ARELLANO, 2004, p.16).

Com a aplicação de um modelo voltado para o gerenciamento do ciclo de vida dos objetos digitais, pode-se dizer que avançamos algumas etapas com relação às políticas de preservação digital, haja vista que as estruturas-base de uma CD "encapsulam" estratégias de conservação que podem, ao longo do tempo, ajudar na definição de padrões que facilitem a vida dos produtores de dados e informações. A utilização – por parte das CD –, de padrões de representação de dados, como os definidos pelo World Wide Web Consortium[6] (W3C), são boas apostas neste sentido.

2.1.3 Áreas de Aplicação

A grande enxurrada de dados científicos, biomédicos e de pesquisas em engenharia gerados em volume cada vez maior são um desafio para a CD. Muitas agências financiadoras de projetos estão comprometidas com estratégias de longo prazo para a provisão de recursos de dados e o desenvolvimento de políticas para o gerenciamento destes dados (KIM, 2014). Entre as diversas esferas de aplicação que têm atualmente investido na proposta de CD, podemos destacar duas: a Educação e a Pesquisa Científica.

A área da Educação, por exemplo, possui diversos trabalhos com este objeto de pesquisa. Segundo Deschaine e Sharma (2015), apesar da ideia de CD ser recente, na verdade educadores têm exercido o papel de curadores digitais já há algum tempo. Professores universitários, por exemplo, exercem tanto o papel de coletores de conteúdo para suas disciplinas quanto o de transmissores de conhecimento acadêmico para a geração atual e as gerações futuras. Os autores complementam dizendo que os professores são confrontados com uma investida de materiais e recursos digitais que tem o potencial de melhorar dramaticamente o processo de ensino e aprendizagem,

[6] https://www.w3.org/

desde que sejam intencionalmente armazenados e processados. Com este propósito em mente, estes materiais digitais deveriam então ser cuidadosamente minerados, organizados e arquivados apropriadamente. Para eles, este processo é a Curadoria Digital (DESCHAINE; SHARMA, 2015).

Quanto à Pesquisa Científica, para Sayão e Sales (2013), além da preocupação das agências financiadoras em garantir retorno dos investimentos, há uma crescente preocupação com a possibilidade de reuso dos dados por pesquisas subsequentes.

> [...] soma-se agora a preocupação com a capacidade de reuso dos dados em outros domínios disciplinares diferentes daqueles para os quais eles foram originalmente gerados. [...] Pelo lado mais pragmático e operacional, um conjunto de atividades gerenciais, técnicas e informacionais fortemente padronizadas - chamado coletivamente de curadoria de dados de pesquisa -, permite que os dados possam ser tratados, arquivados em ambientes digitais confiáveis, preservados e reconfigurados de forma que possam ser aplicados em novos contextos científicos; sirvam de base para novas pesquisas; sejam aproveitados para fins educacionais; e, sobretudo, colaborem para minimizar a duplicação de esforços nas estratégias de criação de dados (SAYÃO; SALES, 2013, p.4).

De maneira geral, a ideia de CD se estrutura como um campo multidisciplinar, que envolve diversas áreas do conhecimento, entre as quais podemos destacar a Ciência da Informação, a Biblioteconomia, a Arquivologia, a Museologia e a Ciência da Computação. Para Dobreva e Duff (2015), entretanto, a comunicação e a interação entre as diferentes áreas que compõem a CD continuam a ser um desafio a ser superado.

Neste trabalho, nos propomos colaborar no sentido de tentarmos expandir esta fronteira, analisando as CD a partir de uma perspectiva Big Data.

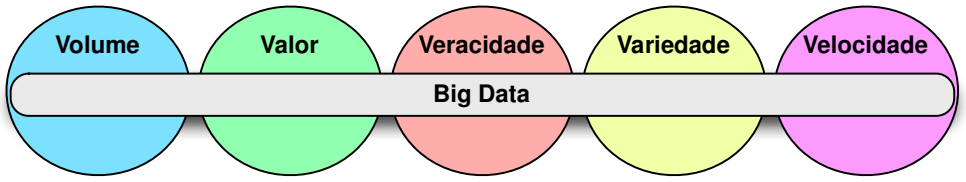
2.2 Big Data

Nos últimos 20 anos, vimos a área de tecnologias da informação e comunicação (TIC) amadurecer e se tornar um dos principais pilares dos negócios em todo o mundo. Atualmente, a extensa maioria das atividades

econômicas e os negócios são a ela associados, utilizando a tecnologia como base de suas operações. Uma problemática vertente, que veio à tona nos últimos anos, diz respeito à quantidade de dados gerados, por estas múltiplas atividades, que têm gerado uma massa de dados com grande volume e de difícil gerenciamento. Há uma frase, que define bem porque a crescente geração de dados dos últimos anos é tão importante para todas as empresas e organizações que estão trabalhando com informação: "você não consegue gerenciar o que você não consegue medir". Esta frase é atribuída a W. Edwards Deming e Peter Drucker (MCAFEE *et al*., 2012).

Hoje, as organizações estão gerando e capturando trilhões de *bytes* de informações dos seus consumidores, fornecedores, operações e de milhões de sensores de redes embutidos em dispositivos do mundo real (celulares, automóveis, etc). Percebe-se uma tendência desses *sites* e serviços com conteúdos multimídia, bem como de pessoas com dispositivos móveis e acesso a ferramentas colaborativas da Web 2.0, de continuarem abastecendo este crescimento exponencial. A estes grandes conjuntos de dados que podem ser capturados, transferidos, agregados, armazenados e analisados, dá-se o nome de Big Data (MANYIKA *et al*., 2011). Big Data está predominantemente associado a duas ideias: armazenamento (*data storage*) e análise de dados (*data analytics*) (WARD; BARKER, 2013).

Figura 1. Características Big Data.



Fonte: Os autores.

Laney (2001) em um relatório técnico do Gartner Group trata da divisão do Big Data em 3 Vs, sendo eles: Volume, Velocidade e Variedade. No relatório, pontos técnicos são abordados, relacionados ao tamanho dos dados gerados, às taxas de crescimento em que estes estão sendo produzidos e ao aumento dos

tipos, representações e formatos de dados que estão surgindo. Esta definição do Gartner foi revisada por alguns trabalhos, inclusive por Beyer e Laney (2012) e NIST (2016), onde foi inserido um quarto V, sendo este de Veracidade. Depois, outros autores estenderam o conceito para um quinto V, sendo este de Valor. Este V é reconhecidamente de difícil caracterização, mas mostra-se como um dos pontos-chave para a agregação de valor em ambientes de CD. Na Figura 1, é possível visualizar a ideia de Big Data está inserida neste contexto.

A seguir, serão apresentados os 5 V's, sendo eles: volume, velocidade, variedade, veracidade e valor.

2.2.1 Volume

A questão do volume em cenários Big Data destaca-se como um dos principais desafios a serem enfrentados por pesquisadores e organizações que desejam gerenciar e extrair valor destas grandes massas de dados. Entretanto, não há uma definição plenamente aceita de como este volume pode ser caracterizado, em termos de tamanho de *datasets*. Em um levantamento empreendido pela empresa Intel (INTEL, 2016), são consideradas cenários Big Data organizações que gerem em média 300 TB (*terabytes)* de dados por semana. No entanto, esta não é uma definição consensual, mas sim, fruto de um estudo com empresas de grande porte, que atuam no ramo de serviços na Internet. Certamente, o número provido pelo estudo da Intel impressiona, mas como citado, não há um número de *megabytes*, *gigabytes* ou *terabytes* que consiga caracterizar se o cenário é Big Data ou não. Por outro lado, há cenários menores tão complexos quanto os citados, que geram, armazenam e manipulam tantos dados, que estas tarefas também se tornam um desafio. Logo, a escala do serviço oferecido e a quantidade de recursos informacionais envolvidos não podem ser as únicas vertentes de classificação.

Para auxiliar no processo de gerenciamento destas massas de dados e informações, várias tecnologias foram desenvolvidas. Saiu-se de um paradigma quase que exclusivamente baseado em bancos de dados relacionais (colunas e tabelas), dos Sistemas Gerenciadores de Bancos de Dados Relacionais – onde os recursos informacionais eram representados sob a forma de dados

estruturados –, para modelos que contemplam dados não-estruturados e semiestruturados, oriundos de várias fontes, de variados formatos, com diferentes demandas, e que necessitam de tecnologias e técnicas diferentes para executar seu gerenciamento. Podemos citar como exemplos dessas técnicas: Business Intelligence (BI), Data Mart, Data Warehouse, ETL (*Extract, Transform and Load*), NoSQL, entre outras. Como exemplos de aplicações tecnológicas já desenvolvidas para o armazenamento e manipulação de recursos informacionais não-estruturados e semiestruturados, podemos citar: Cassandra, R, Hadoop, HBase, Dynamo, Big Table, Redis, entre outros.

2.2.2 Velocidade

Big Data consiste primariamente de um conjunto de dados, caracterizados pelo volume, variedade e velocidade, que necessitam de arquiteturas escaláveis para que os processos de armazenamento, manipulação e análise sejam eficientes (NIST, 2016). Neste sentido, este V trata da problemática de cenários computacionais complexos, com uma quantidade massiva de dados espalhados versus a necessidade de se ter informações em tempo-real, para auxiliar na tomada de decisão, ou ainda, para satisfazer qualquer necessidade informacional.

Mesmo com o avanço das TIC nos últimos anos, ainda é complexa a tarefa de se fazer transitar centenas de *gigabytes* pela Internet. À medida que novas tecnologias de transmissão foram sendo criadas e desenvolvidas, os protocolos clássicos de comunicação entre sistemas não acompanharam o mesmo ritmo. Um exemplo disso é a troca do protocolo de endereçamento de máquinas na Internet, o IPv4 (que possui capacidade de trabalhar com aproximadamente 2^{32} ou 4,3 bilhões de endereços IP) para a sua nova versão, o IPv6[7] (que possui capacidade de endereçar 2^{128} IPs, ou $7{,}9 \times 10^{28}$ mais endereços que o IPV4). O desenvolvimento do IPv6 teve seu início em 1994, somente foi oficializado em 2012, e está caminhando a passos lentos em sua

[7] https://en.wikipedia.org/wiki/IPv6

implantação de fato na Internet. É preciso que se diga que mudanças complexas como esta são críticas e devem ser feitas com o máximo cuidado.

Entretanto, a partir deste exemplo, podemos perceber como alguns novos protocolos de comunicação estão sendo vagarosamente inseridos no contexto da Internet. Isso não impede, no entanto, que os ambientes e aplicações Big Data sigam ganhando complexidade dia a dia.

2.2.3 Variedade

A variedade de tipos de dados – que podem estar representados em diversos formatos, inseridos em múltiplos tipos de repositórios e em domínios totalmente distintos –, torna esta caraterística de extrema importância no contexto das CD.

Atualmente, há uma ampla gama de tipos e formatos de dados, o que torna esta característica bastante complexa. Exemplos destes podem ser arquivos do tipo: *doc*, *docx*, *xls*, *xlsx*, *ppt*, *pptx*, *odt*, *odp*, *pdf*, *jpg*, *jpge*, *png*, entre outros. Ainda dentro desta característica, há outra ampla gama de repositórios estruturados (minoria), semiestruturados e não-estruturados (maioria) existentes. Cada um destes repositórios pode usar tecnologias distintas para preservar digitalmente seus dados, logo, surge um problema de integração ou interoperabilidade destes dados. Podemos citar como exemplos de algumas destas tecnologias: Microsoft SQL Server, MySQL, PostgreSQL, Oracle, Hadoop, Cassandra, MongoDB, Redis, Dynamo, entre outros.

Por fim, há a questão dos múltiplos tipos de domínios existentes. Com o crescimento informacional dos últimos anos, os domínios de conhecimento foram expandidos e diversificados, logo, existem múltiplos tipos de domínios que, de certa forma, podem ser similares e conter interseções entre eles. É verdade também que isto é possível devido à grande quantidade de novas informações geradas sobre determinados campos de conhecimento que, desta forma, contribuem para o processo de aumento da complexidade informacional como um todo.

De qualquer forma, a variedade dos múltiplos tipos de domínios está

relacionada a múltiplos tipos de repositórios digitais e à massiva quantidade de tipos e formatos de dados existentes nestes locais.

2.2.4 Veracidade

A característica da veracidade em cenários Big Data está relacionada à necessidade de se ter um grau de certeza sobre a confiabilidade e a consistência dos dados que se possui, bem como com relação a autenticidade e precisão dos dados. A importância deste V vai ao encontro da massiva quantidade de dados (volume) nestes cenários, bem como dos dados de diversas fontes e tipos (variedade) que precisam ser recuperados no tempo adequado (velocidade). A veracidade exige, no entanto, que estes dados recuperados possuam consistência, precisão, completeza, confiabilidade e autenticidade.

Quando se leva em consideração que, em muitos casos, estes cenários auxiliarão na tomada de decisão de gestores e profissionais da informação, fica evidente que dados imprecisos poderão causar perdas e problemas dos mais variados tipos. Neste caso, não há espaço para informações parciais, inexatas e sem confirmação de origem. Esta característica é complexa de ser efetivamente alcançada num ambiente Big Data, sendo necessário se dispor de um bom arcabouço tecnológico da área de segurança computacional, no qual serão aplicados métodos específicos para garantir a autenticidade e confiabilidade das informações. Além disto, há a necessidade de um esforço adicional de avaliação sobre a qualidade dos dados que serão recuperados, no sentido de se verificar se eles atendem às necessidades informacionais a que se propõem.

2.2.5 Valor

Este V trata sobre como podemos extrair valor do volume de dados, de variados tipos e fontes, para o cenário em questão. Tanto em centros curadores de bibliotecas, quanto de museus, o conceito de valor agregado se aplica sobretudo a coleções temáticas construídas: (i) a partir de objetivos físicos (o somatório sendo maior que as partes); (ii) a partir da documentação que

acompanha os objetos individuais e coleções que proveem o contexto relevante e histórico para a pesquisa, aprendizado e descoberta; e (iii) a partir das habilidades, competência de domínio, e conhecimento dos profissionais curadores das coleções (BEAGRIE, 2006).

Embora nem toda informação digital gerada possua valor de longo prazo, frequentemente, ao menos uma parte significante dela possui. O prazo e a quantidade destas partes vão variar de acordo com as áreas de conhecimento e diferentes categorias nas quais o material se insere. Portanto, o processo de seleção de materiais para serem curados e preservados a longo prazo pode se tornar bastante complexo (BEAGRIE, 2006). Técnicas de Big Data podem auxiliar no processo de identificação e seleção de objetos com maior valor agregado para que o avaliador possa ter informações mais acuradas para poder tomar a decisão de preservar ou não determinado objeto.

2.3 Curadoria digital e Big Data

A relação entre a Curadoria Digital e aplicações Big Data pode ser muito proveitosa, visto a quantidade de dados, informações e conhecimento encapsulados em vários formatos, distribuídos em múltiplos repositórios, coletados em variados domínios. Cria-se, desta forma, um ambiente complexo e de difícil gerenciamento. Juntando este fato à necessidade de preservação digital de recursos informacionais das CD, gera-se uma demanda por melhores métodos e técnicas para busca, tratamento e seleção de novos recursos (futuros objetos digitais nos repositórios das CD) a serem preservados. Fica claro, portanto, a relação próxima e a utilização de técnicas e ferramentas Big Data para fornecimento de recursos informacionais para as CD.

Apesar de tudo isto, atualmente pouco se encontra na literatura técnico-científica sobre novas abordagens que proponham e explicitem esta interseção CD-Big Data. Desta forma, esta se apresenta como um problema em aberto na literatura científica. Importante frisar que a utilização de Big Data para CD não é de propósito geral, visto a complexidade e, por conseguinte, o custo deste tipo de abordagem. Indica-se esta investida quando a escala do cenário a ser

implantado justificar o investimento no mesmo.

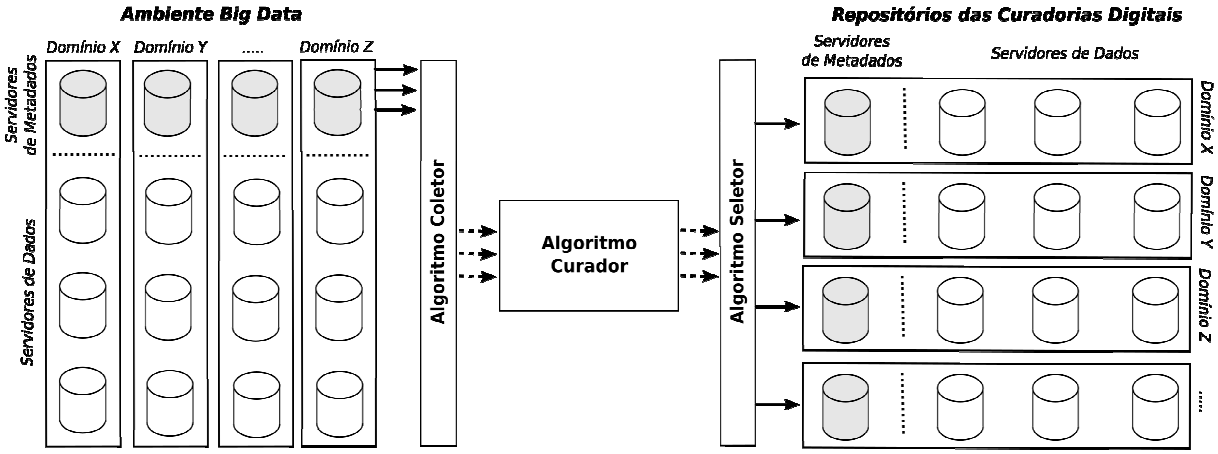
O ciclo de vida da ciência dos dados para Big Data consiste em quatro estágios, sendo eles: (i) coleção; (ii) preparação; (iii) análise; e (iv) ação. A coleção é o estágio que seleciona os dados crus (*raw data*) e os armazena no repositório. A preparação envolve um conjunto de processos e técnicas que processam os dados crus, organizando e limpando os mesmos. Na fase de análise, as técnicas de produção e sintetização de conhecimento são aplicadas. E por fim, na ação, o conhecimento sintetizado e processado é utilizado para agregar valor à organização (NIST, 2016).

3 PROPOSTA DE MODELO DE CURADORIA DIGITAL BASEADO EM BIG DATA

Baseado no referencial teórico acima apresentado, este artigo apresenta uma proposta de modelo baseada nas características de Big Data, aplicada no contexto das Curadorias Digitais.

Este modelo tem a função base de realizar a função de mineração de dados e recursos informacionais em bases de dados estruturadas, semiestruturadas e não-estruturadas, selecionando e sugerindo novos objetos digitais aos avaliadores (curadores), que decidirão pela preservação, ou não, destes. O objetivo é auxiliar na busca, seleção e oferta de novos objetos digitais aos curadores (humanos ou agentes de software). Na Figura 2 é possível visualizar a proposta geral do modelo Big Data para suportar um processo de seleção semiautomática nas CD.

Figura 2. Visão Geral do Modelo de Curadoria Digital em Ambiente Big Data



Fonte: Os autores.

O cenário Big Data foi dividido em múltiplos domínios do conhecimento (X, Y, ..., Z), que representam domínios específicos para cada variedade de dados a ser analisada para preservação, tais como: músicas, vídeos, documentos, objetos de aprendizagem, entre outros. Cada um dos domínios está representado com servidores de dados, onde de fato os recursos informacionais ainda não coletados estão armazenados, e servidores de metadados, onde estão os descritores destes recursos. Estes domínios são implementados, na prática, por meio de bancos de dados não-estruturados na Internet, como por exemplo: redes sociais, blogs, sites de notícias, repositórios de conteúdo, entre outros. Ainda, podem ser representados em bancos de dados estruturados, como por exemplo: bases de dados de bibliotecas, acervos on-line, enciclopédias, bases de dados abertas de periódicos, entre outros.

Todos estes múltiplos bancos de dados estão interconectados por meio de um **Algoritmo Coletor**, que exerce o papel de "mediador" (*middleware*[8]), e tem a função de integrar as múltiplas tecnologias e as bases de dados conectadas, auxiliando no processo de limpeza e seleção dos objetos digitais advindos dos múltiplos domínios do conhecimento pesquisados.

[8] Pode ser entendido como um software mediador entre outros componentes de softwares. São comumente utilizados como agentes integradores para troca de dados e interconexão de sistema.

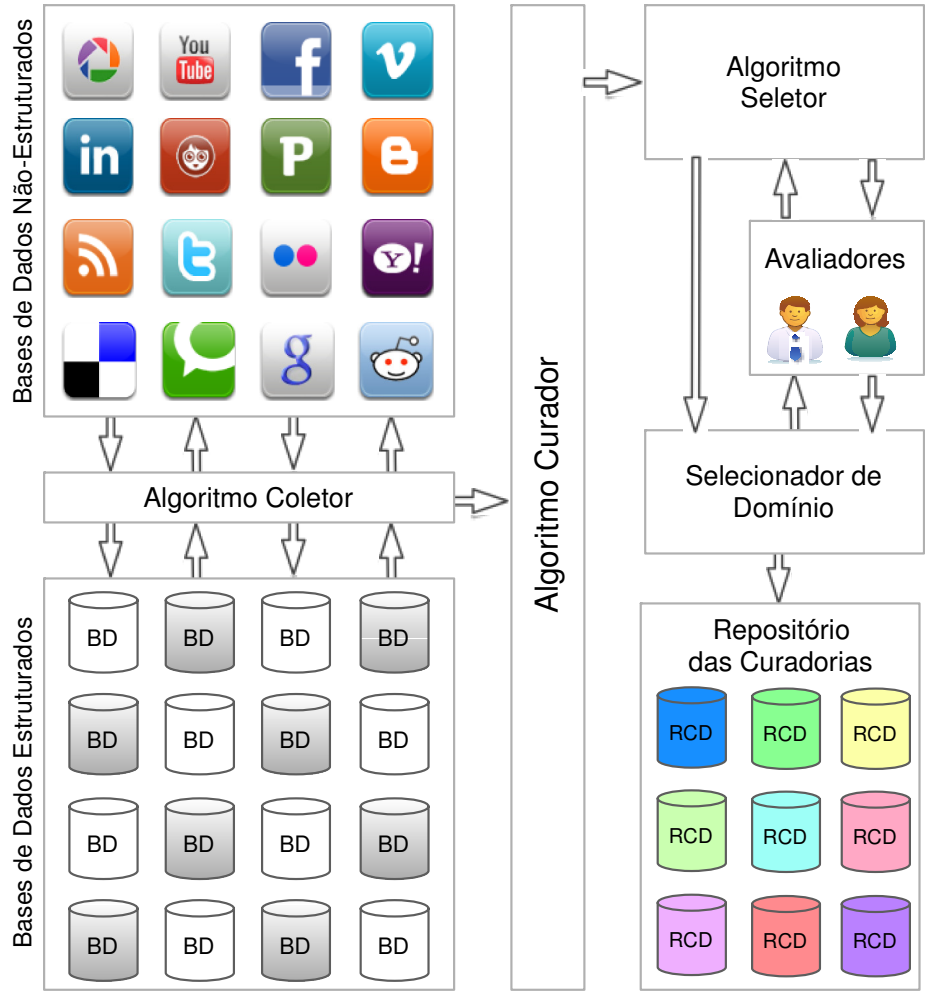
O Algoritmo Coletor é responsável pelo processo de busca, tratamento e seleção dos dados brutos, e depois envia-os para o componente chamado de **Algoritmo Curador**, que tem a função de estabelecer a relevância do recurso para as curadorias que estão interconectadas. No entanto, é necessário que o Algoritmo Coletor efetue uma filtragem preliminar dos recursos informacionais coletados, antes de enviá-los ao Algoritmo Curador, que irá efetuar uma análise mais acurada. Esta filtragem preliminar tem por objetivo reduzir o custo computacional dispendido pelo Algoritmo Curador, que não precisa executar todas as análises heurísticas necessárias num enorme montante de dados crus, o que poderia causar lentidão no processo ou, ainda, inviabilizá-lo por completo.

No Algoritmo Curador estão programadas heurísticas bem definidas, separadas por tipo de domínio de conhecimento (músicas, vídeos, documentos texto, etc.) que auxiliam o agente de software a tomar a decisão sobre se deve ser criado um objeto digital para representar aquele recurso coletado ou não. Em caso positivo, o objeto digital é entregue ao **Algoritmo Seletor**, como um possível objeto a ser preservado pelas CD. A abordagem semiautomática ocorre neste ponto, pois seleção do repositório da CD poderá ser feita por um agente de software ou por um curador humano. De forma prática, quando uma curadoria, de determinado tipo, é integrada ao sistema, é solicitada a criação de heurísticas específicas para ela, que serão integradas aos Algoritmos Curador e Seletor.

Depois de aplicadas as heurísticas pelo Algoritmo Curador, duas ações possíveis são tomadas para cada recurso coletado e filtrado: ou ele é identificado como relevante ou é identificado como descartável. No primeiro caso, o objeto digital correspondente é enviado para um avaliador (humano ou não) que terá a função de moderar se o mesmo é relevante ou não para sua CD. Se ele for aprovado pelo avaliador, os metadados então são extraídos e armazenados nos serviços de metadados e, por fim, o objeto digital é de fato armazenado nos serviços de armazenamento da CD, pelo Algoritmo Seletor. Um importante aspecto desta abordagem é que as CD também são divididas em domínios específicos, para facilitar o cruzamento e gerenciamento dos objetos digitais.

A Figura 3 ilustra mais detalhadamente o processo de semiautomático de seleção de objetos digitais.

Figura 3-Processo Semiautomático de Seleção de Objetos Digitais.



Fonte: Os autores.

O processo de semiautomático de seleção se inicia com a identificação das fontes de dados, sejam elas bancos de dados estruturados, semiestruturados ou não-estruturados. No caso dos bancos de dados estruturados, estes podem ser bases de dados científicas, de periódicos, bancos de dados corporativos, entre outros. No caso dos bancos de dados semiestruturados e não-estruturados, algumas possíveis fontes de dados são: Linkedin, Facebook, Blogger, Vime, Picasa, Youtube, entre outros. Todas estas bases de dados são interconectadas e seus recursos coletados e tratados por

meio de Algoritmos Coletores antes de serem enviados para o Algoritmo Curador, para dar sequência ao processo já explicitado acima.

Para exemplificar um caso de uso que não envolva o curador humano, podemos citar uma CD que tem por função preservar vídeos de jornais noticiosos da televisão. A heurística e as fontes de dados são relativamente bem definidas e, neste caso, provavelmente não haverá a necessidade de avaliação humana anterior ao armazenamento do objeto digital no repositório da CD equivalente.

4 CONSIDERAÇÕES FINAIS

Neste trabalho, foi apresentada uma proposta de modelo para curadorias digitais baseadas em ambientes Big Data, tratando de forma semiautomática o processo de seleção de objetos digitais a serem preservados. Desta forma, propusemos um modelo conceitual multidomínio, que tem como função interagir com múltiplos repositórios e fontes de informação, e, a partir de heurísticas pré-definidas, selecionar objetos digitais que possuam valor para às Curadorias Digitais, facilitando assim, o processo de busca e seleção em grandes massas de dados. Como complemento, também foi apresentado um modelo de operação do processo semiautomático de seleção, que teve o objetivo de facilitar o entendimento do funcionamento do modelo conceitual.

Sobre às perguntas de pesquisa, pode-se dizer sobre a primeira (“É possível utilizar ambientes Big Data como fonte de recursos informacionais para Curadorias Digitais? ”) que sim, é possível. O extenso cenário de ambientes Big Data, se bem utilizado e tratado, apresenta-se como uma fonte rica de recursos informacionais a serem descobertos. Sobre a segunda pergunta de pesquisa (“ Como as técnicas e ferramentas Big Data podem auxiliar no processo de busca e seleção de recursos informacionais para serem preservados em Curadorias Digitais?”), a sugestão deste trabalho é que métodos heurísticos sejam utilizados, em conjunto com algoritmos mediadores, que auxiliem no processo de

coleta e seleção dos recursos informacionais e, ainda, que possam auxiliar no tratamento e encapsulamento dos mesmos em objetos digitais.

A sequência deste trabalho irá focar no desenvolvimento e especificação destes métodos heurísticos, que serão utilizados pelos Algoritmos Coletor, Curador e Seletor. Além disso, o modelo proposto será testado em um cenário computacional real, por meio da utilização de repositórios públicos e privados para a avaliação do comportamento do mesmo. Subsequentemente, será feita a inserção da proposta no processo de ciclo de vida das curadorias digitais.

Para finalizar, é necessário ressaltar que este é um trabalho seminal, que visa diversos desdobramentos. Acreditamos que há espaço para o estudo do ciclo de vida dos dados e o ciclo de vida das curadorias digitais, que podem ser cruzados com o ciclo de vida do Big Data, na tentativa de se encontrar uma integração suave para a criação, alimentação e manutenção dos repositórios resultantes dos processos de curadorias digitais.

REFERÊNCIAS

ABBOT, Daisy. What is digital curation? Edinburgh, UK: Digital Curation Centre, 2008. Disponível em: <http://www.era.lib.ed.ac.uk/bitstream/1842/3362/3/Abbott%20What%20is%20digitl%20curation_%20_%20Digital%20Curation%20Centre.doc>. Acesso em: 02 set. 2016.

ARELLANO, Miguel Angel. Preservação de documentos digitais. **Ci. Inf.,** v. 33, n. 2, p.15-27, ago. 2004.

BEAGRIE, Neil. Digital Curation for Science, Digital Libraries, and Individuals. **The International Journal Of Digital Curation**, Edinburgh, v. 1, n. 1, p.3-16, out. 2006.

BEYER, M. A.; LANEY, D.. The importance of big data: A definition. 2012. Disponível em: <https://www.gartner.com/doc/2057415/importance-big-data-definition>. Acesso em: 03 set. 2016.

CORRÊA, E. S.; BERTOCCHI, D. A cena cibercultural do jornalismo contemporâneo: Web semântica, algoritmos, aplicativos e curadoria. **MATRIZes**, São Paulo, v. 5, n. 2, p.123-144, jun. 2012.

DALLAS, C. Digital curation beyond the “wild frontier”: a pragmatic approach. **Arch Sci,** p.1-37, set. 2015.

DESCHAINE, M.; SHARMA, S. A. The Five Cs of Digital Curation: Supporting Twenty-First-Century Teaching and Learning. **Insight: A Journal of Scholarly Teaching,** v. 10, n. 1, p.19-24, 2015.

DOBREVA, M.; DUFF, W. M. The ever changing face of digital curation: introduction to the special issue on digital curation. **Arch Sci,** v. 15, n. 2, p.97-100, mar. 2015.

FERREIRA, M. *et al*. **ESTADO DA ARTE EM PRESERVAÇÃO DIGITAL**. Lisboa: Repositório Científico de Acesso Aberto de Portugal, 2012.

HIGGINS, S. Digital Curation: The Emergence of a New Discipline. **International Journal Of Digital Curation,** v. 6, n. 2, p.78-88, out. 2011.

INTEL. **Peer Research Report**: Big Data Analytics. Disponível em: <http://www.intel.co.za/content/www/za/en/big-data/data-insights-peer-research-report.html>. Acesso em: 03 set. 2016.

KIM, J. Growth and trends in digital curation research: The case of the international journal of digital curation. In: PROCEEDINGS OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 2014, Seatle, Estados Unidos. **Proceedings...** Seatle: Asist, 2014. p. 1 - 4.

LANEY, D. 3D data management: Controlling data volume, velocity and variety. META Group Research Note 6, fev. 2001.

MANYIKA, J. *et al*. **Big data:** The next frontier for innovation, competition, and productivity. Mckinsey Global Institute, 2011.

MCAFEE, A.; BRYNJOLFSSON, E.; DAVENPORT, T. H.; PATIL, D. J.; BARTON, D. **Big data: The management revolution**. Harvard Business Review, v. 90, no. 10, p. 61-67, out. 2012.

NIST. **Big Data Definitions**. Disponível em: http://dx.doi.org/10.6028/NIST.SP.1500-1. Acesso em: 03 set. 2016.

SAYÃO, L. F.; SALES, L. F. CURADORIA DIGITAL: um novo patamar para preservação de dados digitais de pesquisa. **Informação e Sociedade: Estudos**, João Pessoa, v. 22, n. 3, p.179-191, set./dez., 2012.

SAYÃO, L.; SALES, L. DADOS DE PESQUISA: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 6, n. 1, dez. 2013.

WARD, J. S.; BARKER A. B Undefined by data: a survey of big data definitions. **arXiv preprint arXiv:1309.5821**, out. 2013.

YAKEL, E. Digital curation. **OCLC Systems & Services**: International digital library perspectives, vol. 23, n. 4, p.335 – 340, abr. 2007.

Title

Digital curation: a proposal of a semi-automatic digital object selection-based model for digital curation in Big Data environments

Abstract
Introduction: This work presents a new approach for Digital Curations from a Big Data perspective.
Objective: The objective is to propose techniques to digital curations for selecting and evaluating digital objects that take into account volume, velocity, variety, reality, and the value of the data collected from multiple knowledge domains.
Methodology: This is an exploratory research of applied nature, which addresses the research problem in a qualitative way. Heuristics allow this semi-automatic process to be done either by human curators or by software agents.
Results: As a result, it was proposed a model for searching, processing, evaluating and selecting digital objects to be processed by digital curations.
Conclusions: It is possible to use Big Data environments as a source of information resources for Digital Curation; besides, Big Data techniques and tools can support the search and selection process of information resources by Digital Curations.

Keywords: Digital Curation, Big Data, Digital Objects.

Título

Curación de contenidos digitales: propuesta de un modelo para curación de contenidos digitales en ambientes Big Data basado en un enfoque semiautomático de selección de objetos digitales

Resumen
Introducción: Propone un nuevo abordaje en la curación de contenidos digitales a partir de una perspectiva de Big Data.
Objetivo: El objetivo de este trabajo es proponer técnicas de selección y evaluación de objetos para curación de contenidos digitales, teniendo en cuenta el volumen, velocidad, variedad, veracidad y valor de los datos generados en múltiples dominios del conocimiento.

Metodología: Es una investigación exploratoria de naturaleza aplicada, con abordaje cualitativo. A partir de la heurística, este proceso semiautomático puede ser llevado a cabo tanto por curadores humanos como por agentes de software.
Resultados: Se propone un modelo para búsqueda, tratamiento, evaluación y selección de objetos digitales para ser tratados en la curación de contenidos.
Conclusiones: Se concluye que es posible utilizar Big Data como fuente de recursos de información para curación de contenidos, y que algunas técnicas y herramientas de Big Data pueden ayudar en el proceso de búsqueda y selección de recursos de información para ser preservados en bases de datos.

Palabras clave: Curación de contenidos, Big Data, Objetos Digitales.

Enviado em: 17.07.2016.
Aceito em: 20.11.2016.