Uma proposta de taxonomia para dados de pesquisa

Luana Farias Sales[1], Luís Fernando Sayão[2]

1 0000-0002-3614-2356 + IBICT; Rio de Janeiro, Brasil. luanafsales@gmail.com
2 0000-0002-6970-0553 + CNEN; Rio de Janeiro, Brasil. lsayao@cnen.gov.br

Tipo de trabalho: Comunicação

Palavras-chave: Dados de pesquisa; taxonomía; tipos de dados

RESUMO
No contexto da ciência contemporânea, os dados de pesquisa deixam de ser meros subprodutos das atividades de pesquisa e ressurgem como protagonistas na busca por novos conhecimentos. Esse fenômeno é impulsionado pelas tecnologias digitais que criam as condições para o surgimento de um genuíno *big* data, científico e engendram processos de pesquisa baseados na coleta, geração, processamento e análise de massivas quantidades de dados estruturados em bases de dados. Os pesquisadores, instituições acadêmicas, formuladores de políticas científicas e agências de fomento começam a compreender que os dados de pesquisa se bem gerenciados se tornam recursos informacionais imprescindíveis que podem ser compartilhados e reusados como *input* para novas pesquisas. Entretanto, os dados, diferentes das publicações, são heterogêneos, diversificados, gerados para diferentes propósitos, por diferentes tecnologias e em domínios disciplinares específicos. Observa-se que há uma lacuna terminológica que dificulta a gestão desses ativos informacionais. Partindo desse ponto, o presente trabalho, aceita o desafio de propor uma taxonomia para a classificação de tipos de dados de pesquisa, ancorado na abordagem teórico-metodológico da Organização do Conhecimento.

1 Apresentação

Cientistas de todo o mundo têm abordado a necessidade de aumentar o acesso global aos dados de pesquisa que são produzidos em quantidade cada vez maior. Isto acontece essencialmente devido à tecnologia digital que se torna cada vez mais um elemento onipresente nos processos da construção do conhecimento científico e permite também que esse conhecimento seja compartilhado e construído de forma cooperativa. Os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a entender que estes dados, se preservados e bem gerenciados, constituem uma excelente fonte de recursos informacionais que podem ser compartilhados e reutilizados como insumo para novas pesquisas.

A Declaração de Berlim sobre o Acesso Aberto ao Conhecimento em Ciências e Humanidades, publicada em 2003, amplia o escopo do que se entende por acesso livre ao definir que as "contribuições de acesso livre incluem resultados de pesquisas científicas originais, dados não processados e metadados, fontes originais, representações digitais de materiais pictóricos e gráficos e materiais acadêmicos multimídia" (BERLIN, 2003).

Como um ponto de inflexão histórica da ascensão dos dados , a *D-Lib Magazine*[1] – um periódico importante no que envolve as pesquisas em bibliotecas digitais – publicou no início de 2011, um número especial sobre dados de pesquisa onde estão apresentadas questões como acesso livre, curadoria digital, aquisição e gestão, qualidade e confiabilidade e as possíveis conexões entre dados de pesquisa e as publicações acadêmicas tradicionais. Em 2014, essa mesma revista volta a publicar outro número sobre dados de pesquisa, mas dessa vez, enfatizando a criação do *Research Data Alliance*, também conhecido como RDA – uma aliança entre interessados na preservação e no tratamento de dados de pesquisa.

Assim, considerando a relevância da temática "dados de pesquisa" para o mundo científico, bem como temas correlatos, como Gestão de dados, Curadoria Digital, Ciência aberta, e-Science etc. e ainda considerando a interdisciplinaridade da própria da natureza do objeto "dado de pesquisa" que pode se originar em diferentes áreas, com metodologias, jargões e práticas próprias de cada domínio, o presente trabalho se coloca diante de uma lacuna que é a ausência de um instrumento terminológico e classificatório que possibilite a gestão e a curadoria de dados pesquisa, bem como a interlocução entre os atores que pesquisam esta temática e estabelecem políticas, normas e padrões para propiciar o uso e o reuso de dados em seu potencial máximo por diversas disciplinas.

Plataformas de gestão de dados de pesquisa, bem como políticas de gestão necessitam de uma categorização precisa dos tipos de dados que serão gerenciados. Neste sentido, sob a abordagem teórico-metodológica da Organização do Conhecimento, o presente trabalho vem propor uma taxonomia para classificação de tipos dados de pesquisa.

2 Dado de pesquisa: uma tentativa de definição

Inspirado na Teoria do Conceito de Dahlberg (1978) e na Teoria de Eugene Wüster (1981), o presente estudo, parte de uma tentativa de compreensão do conceito de dado de pesquisa. Em uma abordagem exploratória, encontramos algumas definições que mereceram ser destacadas aqui.

O National Research Council dos EUA, em seu relatório sobre direitos privados e interesse público em bases de dados técnico-científicas, define dados como "fatos, números, letras, símbolos que descrevem um objeto, uma condição, uma situação ou outro fator" (NATIONAL RESEARCH COUNCIL, 1999, p.15). Esta definição pode variar consideravelmente entre colaboradores e de acordo com a área em que são utilizados.

A Organização para Cooperação e Desenvolvimento Econômico (OCDE, 2007), em seu guia para acesso aos dados de pesquisas financiadas por recursos públicos, define como dados de pesquisa "registros de fatos usados como fontes primárias na investigação científica e que geralmente são aceitos na comunidade científica como necessários para a validação dos resultados da pesquisa."

A Secretaria de Gestão e Orçamento dos EUA define dados de pesquisa como aqueles "coletados, observados ou criados para fins de análise para produzir resultados originais de pesquisa". A Universidade de Edmburg os define como "... material factual registrado comumente aceito na comunidade científica como necessário para validar os resultados da pesquisa ..." Já o National Endowment for the Humanities define os dados como

[1] Disponível em: <http://www.dlib.org/dlib/january11/01contents.html>. *Acesso em: 20 maio 2013*. O periódico encerrou suas atividades em julho de 2017.

“materiais gerados ou coletados durante a realização de pesquisas ...". Como pode-se perceber são inúmeras as definições de dados de pesquisa, Borgman (2011) explica o porquê da dificuldade em definir dados de pesquisa. Para ela “Informação é um conceito complexo com centenas de definições [...]. Dado [por sua vez] é um conceito simples com poucas definições, porém sujeito a muitas e diferentes interpretações”. Ou seja, o que dificulta atribuir uma definição consensual ao dado de pesquisa é o fato idiossincrático que ele pode ser muitas coisas diferentes para pessoas e circunstâncias diferentes, ou conforme Buckland (1991) “os dados existem apenas aos olhos do observador”. Isto acontece porque dado de pesquisa é dependente de interpretação. O pesquisador é um interpretador de dados científicos.

Neste sentido, uma sequência de bits proveniente de um sensor sísmico é dado de pesquisa para os sismólogos; amostras de rochas são dados de pesquisa para um geomorfologista; conversas gravadas são dados de pesquisa para sociólogos; e inscrições em cuneiformes são dados de pesquisa para quem estuda linguagens do Oriente. “Porém, os cuneiformes podem ser também dados para o arqueólogo ou para o ambientalista que buscam padrões climáticos históricos; de forma similar, os dados sísmicos podem ser úteis para biólogos que estudam comportamento animal”. (BORGMAN, 2007, p.119)

Nesta pesquisa propomos a seguinte definição:

Dado de pesquisa é todo e qualquer tipo de registro coletado, observado, gerado ou usado pela pesquisa científica, tratado e aceito como necessário para validar os resultados da pesquisa pela comunidade científica.

Essa definição é suficientemente ampla para abarcar todas as possibilidades de dados de pesquisa. No entanto, é importante destacar que para que o registro se configure como dado de pesquisa, ele precisa ser tratado e aceito pela comunidade. Isso coloca em voga a obrigação de uma gestão mínima, com atribuição de metadados que tornem o registro compreensível à comunidade científica.

De acordo com Borgman (2010, p.3), alguns tipos de dados têm tanto valor imediato e duradouro, alguns ganham valor ao longo do tempo, outros têm valor transiente, alguns dados são capturados num momento específico e são se repetirão jamais, enquanto outros são passíveis de serem reproduzidos. Tipos de dados podem incluir, por exemplo, números, imagens, textos, vídeos, áudio, *software*, algoritmos, equações, animações, modelos, simulações. Além do mais, a noção de dados pode variar consideravelmente entre pesquisadores e, ainda mais, entre áreas do conhecimento. A constatação de que os dados são gerados para diferentes propósitos, por diferentes comunidades acadêmicas e científicas e por meio de diferentes processos intensifica ainda mais essa percepção de diversidade e torna imperativo o estabelecimento de uma taxonomia de tipos de dados de pesquisa.

Essa heterogeneidade intrínseca aos dados de pesquisa implica na necessidade de formular estratégias de gestão de amplo espectro que englobem os vários tipos de dados. O reconhecimento dessas diferenças torna-se crucial para diversas ações no escopo da gestão de dados de pesquisa e do ciclo de vida da curadoria, como por exemplo, para as fases de desenvolvimento de coleções de dados, fase de preservação, fase de versionamento e até mesmo para o reuso. Assim, o presente trabalho vem apresentar uma proposta de taxonomia para dados de pesquisa que tem por finalidade auxiliar o gestor

dos dados, seja o pesquisador, ou o bibliotecário no seu papel de curador a desempenhar suas tarefas de forma mais efetiva.

3 Material e Método

A pesquisa realizada se iniciou a partir de uma abordagem exploratória e de cunho qualitativo que se configurou como um levantamento na literatura dos conceitos e tipos de dados de pesquisa. Em seguida, a pesquisa partiu para uma abordagem empírica que se configurou no levantamento dos tipos de dados existentes em uma área do conhecimento multidisciplinar: a área de ciências nucleares. Essa abordagem empírica teve por finalidade verificar a ausência de algum tipo de dado. O levantamento foi realizado a partir de entrevista com os líderes de pesquisa de uma instituição brasileira voltada a pesquisa nessa área: o Instituto de Engenharia Nuclear da Comissão Nacional de Energia Nuclear.

4 Resultados: uma classificação para dados de pesquisa

No levantamento realizado algumas tentativas de classificação dos dados foram encontradas, como a da National Science Foundation (NSF) (2007), a de Borgman (2010), de Harvey (2010), de Lyon (2007) e da OCDE (2007) e, em uma tentativa de sistematização, chegou-se ao seguinte quadro:

Quadro 1: Tentativa de sistematização a partir da literatura

CARACTERÍSTICA DE DIVISÃO	NSF	BORGMAN	HARVEY	LYON
Quanto à natureza	Número Imagem Software			
Quanto à origem	Observacionais Computacionais Experimentais	Registro		
Quanto ao nível de processamento	Intermediário Finais			
Quanto ao estágio de geração		Brutos Derivados		Crus Primários Derivados
Quanto à formação de coleções			De pesquisa De comunidade De referência	
Quanto à mutabilidade				Canônicos Episódicos

Fonte: Os autores

No entanto, essas classificações eram limitadas a uma visão única sobre os dados. Juntar essas múltiplas visões parecia ser uma saída, mas elas se sobrepunham em alguns conceitos e pareciam divergir-se em outros. Além disso, com a abordagem empírica da pesquisa, que se deu a partir da entrevista realizada com pesquisadores da área de ciências nucleares, percebeu-se a ausência de alguns tipos de dados, como era pressuposto. Com os dados identificados, foram levantadas suas definições na literatura. Assim, chegou se no quadro a seguir:

Quadro 2: Definição dos tipos de dados e encontrados na área de Ciências Nucleares

TIPOS DE DADOS		DEFINIÇÃO
Número	Medidas	Quantidade fixada por um padrão para determinar as dimensões ou o valor de uma grandeza da mesma espécie
	Resultados de levantamentos	Resultado de pesquisa prévia e mais ou menos aprofundada de um fenômeno, antes de se fazer um projeto, um programa, uma pesquisa científica etc. (coleta)
	Resultados de experimentos	Resultado de trabalho científico que se destina a verificar um fenômeno.
	Fórmulas	Expressão concisa e rigorosa, constituída em geral de símbolos, que resumeum certo número de dados
	Equações	Redução de uma questão, um problema intrincado, a pontos simples e claros, para facilitar a obtenção de uma solução
	Algoritmos	Seqüência finita de regras, raciocínios ou operações que, aplicada a um número finito de dados, permite solucionar classes semelhantes de problemas
Multimídia	Imagens	Representação da forma ou do aspecto de ser ou objeto por meios artísticos
	Vídeo	Técnica de reprodução eletrônica de imagens em movimento
	Áudio	Sinal sonoro; som
	Animações	Ato ou efeito de imprimir movimento ou aceleração
	Filme	Seqüência de imagens registradas em filme cinematográfico ou videoteipe, para exibição em movimento ou não;
	Fotografia	Imagem obtida por arte ou processo de reprodução sobre uma superfície fotossensível (como um filme), pela ação de energia radiante, esp. a luz
Software	Bases de dados	Conjunto de dados inter-relacionados sobre determinado assunto, armazenados em sistemas de processamento de dados segundo critérios preestabelecidos (reúne)
	Simulações	Teste, experiência ou ensaio em que se empregam modelos para simular o ser humano, em especial em casos de grande perigo de vida
	Códigos nucleares	Programa de computador que representam as simulações matemáticas do núcleo do reator.
Visualização	Tabelas	Quadro sistemático de consulta de dados
	Gráficos	Curva num sistema de coordenadas, que representa uma função [A curva pode ser substituída por uma superfície, uma série de colunas etc.]
	Diagramas	Representação gráfica, por meio de figuras geométricas (pontos, linhas, áreas etc.), de fatos, fenômenos, grandezas, ou das relações entre eles
	Modelos em 3D	Modelo em formato tridimensional, que inclui a idéia de profundidade
	Modelos reduzidos	Esquema que possibilita a representação de um fenômeno ou conjunto de fenômenos físicos e eventualmente a previsão de novos fenômenos ou propriedades, tomando como base um certo número de leis físicas, em geral obtidas ou testadas experimentalmente
	Desenhos	Representação de seres, objetos, idéias, sensações, feita sobre uma superfície, por meios gráficos, com instrumentos apropriados

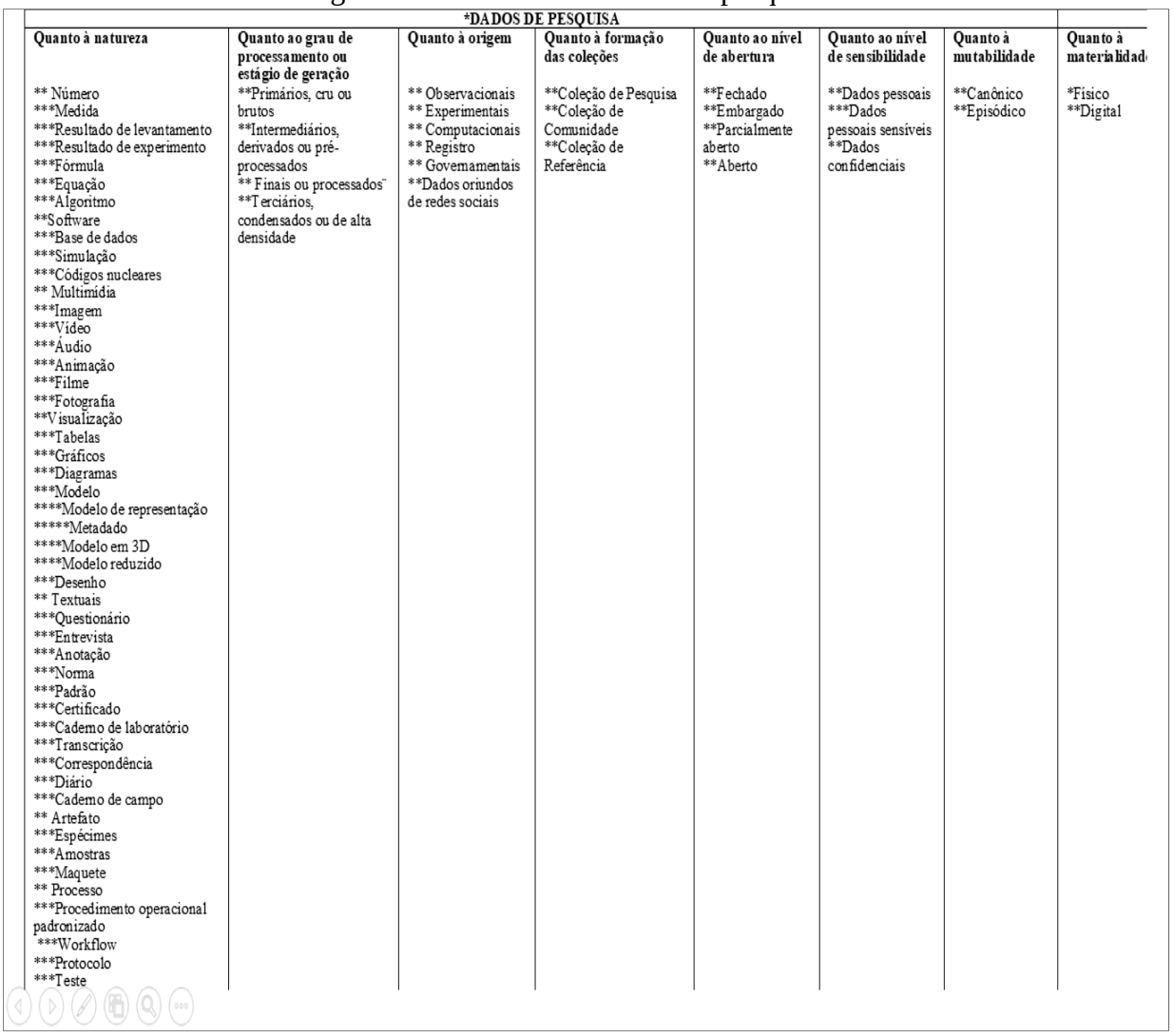
Textuais	Metadados	Dados que registram e preservam dados
	Questionários	Sequência de perguntas feitas para servir de guia a uma investigação
	Entrevistas	Coleta de declarações tomadas para divulgação
	Anotações	Indicação escrita breve
	Normas	Aquilo que regula procedimentos ou atos;
	Padrões	Base de comparação, algo que o consenso geral ou um determinado órgão oficial consagrou como um modelo aprovado. objeto que serve de modelo para outro
	Certificados	Documento no qual se atesta a existência de certo fato e dele se dá ciência
	Caderno de laboratório	Ferramenta usada por pesquisadoresde várias áreas para fazer anotações sobre a pesquisa quando executada em laboratórios.
	Transcrição	Passar para o papel ou equivalente (algo) que está sendo ouvido (p.ex., um texto de discurso, uma música etc.)
	Correspondências	Intercâmbio de mensagens, cartas etc. entre pessoas, promovido através de serviço próprio
	Diário	Escrito em que se registram os acontecimentos de cada dia
	Caderno de campo	Ferramenta usada por pesquisadoresde várias áreas para fazer anotações quando executam trabalhos de campo. É um exemplo clássico de fonte primária.
Artefatos	Espécimes	Exemplo, amostra, modelo
	Amostras	Pequena porção de alguma coisa dada para ver, provar ou analisar, a fim de que a qualidade do todo possa ser avaliada ou julgada
	Maquete	Representação em escala reduzida de uma obra de arquitetura ou engenharia a ser executada. Cenário em miniatura destinado a filmagens de estúdio, quando a obtenção de certas imagens, em ambientes ou paisagens reais, se torna muito difícil ou impraticável; reprodução em miniatura de edifícios, meios de transporte, paisagens etc., us. na simulação de peripécias impossíveis de filmar (p.ex., cenas de catástrofes)
Processos	Procedimentos operacionais padronizados	Procedimento que busca fazer com que um processo, independente da área, possa ser realizado sempre de uma mesma forma, permitindo a verificação de cada uma de suas etapas. Ele deve ser escrito de forma detalhada para a obtenção de uniformidade de uma rotina operacional, seja ela na produção ou na prestação de serviços.
	Workflows	Sequência de passos necessários para que se possa atingir a automação de processos de negócio, de acordo com um conjunto de regras definidas, envolvendo a noção de processos, permitindo que estes possam ser transmitidos de uma pessoa para outra de acordo com algumas regras.
	Protocolos	Planejamento que visa responder uma pergunta ou problema em evidência, definindo a estrutura da pesquisa, selecionando o tipo e o número de variáveis a serem estudadas, e analisando os resultados encontrados
	Teste	Exame crítico ou prova das qualidades de uma pessoa ou coisa
Outros	Phanton ou Manequim	UP Simulador de Tecido Material que possui as mesmas características que o tecido humano com relação à absorção e espalhamento da radiação ionizante.

Fonte: Os autores

Posteriormente, após o estudo conceitual do que cada autor compreendia como tipo de dado de pesquisa, chegou-se a uma proposta de sistematização que tentou incluir todas

as visões encontradas na literatura estudada e ainda acrescentando os tipos indicados pelos pesquisadores durante as entrevistas. A sistematização realizada se pautou no método analítico-sintético de Ranganathan (1967), tentando, sempre que possível se valer de seus cânones e princípios para ordenação dos conceitos.

Figura 1: Taxonomia de dados de pesquisa



*DADOS DE PESQUISA							
Quanto à natureza	**Quanto ao grau de processamento ou estágio de geração**	**Quanto à origem**	**Quanto à formação das coleções**	**Quanto ao nível de abertura**	**Quanto ao nível de sensibilidade**	**Quanto à mutabilidade**	**Quanto à materialidad**
** Número ***Medida ***Resultado de levantamento ***Resultado de experimento ***Fórmula ***Equação ***Algoritmo **Software ***Base de dados ***Simulação ***Códigos nucleares ** Multimídia ***Imagem ***Vídeo ***Áudio ***Animação ***Filme ***Fotografia **Visualização ***Tabelas ***Gráficos ***Diagramas ***Modelo ****Modelo de representação *****Metadado ****Modelo em 3D ****Modelo reduzido ***Desenho ** Textuais ***Questionário ***Entrevista ***Anotação ***Norma ***Padrão ***Certificado ***Caderno de laboratório ***Transcrição ***Correspondência ***Diário ***Caderno de campo ** Artefato ***Espécimes ***Amostras ***Maquete ** Processo ***Procedimento operacional padronizado ***Workflow ***Protocolo ***Teste	**Primários, cru ou brutos **Intermediários, derivados ou pré-processados ** Finais ou processados **Terciários, condensados ou de alta densidade	** Observacionais ** Experimentais ** Computacionais ** Registro ** Governamentais **Dados oriundos de redes sociais	**Coleção de Pesquisa **Coleção de Comunidade **Coleção de Referência	**Fechado **Embargado **Parcialmente aberto **Aberto	**Dados pessoais ***Dados pessoais sensíveis **Dados confidenciais	**Canônico **Episódico	*Físico **Digital

Fonte: Os autores

5 Considerações Finais

De acordo com duas premissas fundamentais apresentadas por Souza (2012, p.4) no que diz respeito às classificações "os mesmos objetos e ideias podem ser organizados e representados de formas diferentes e toda classificação está relacionada a um propósito definido de construção e uso de informação". Sendo assim, todas essas classificações refletem o propósito para o qual foram construídas e um contexto específico. Neste sentido, é provável que a taxonomia aqui apresentada ainda deixe de fora algum tipo de dado, principalmente porque o levantamento empírico realizado abrangeu apenas um domínio específico.

Confirmando ainda essa afirmação, Borgman (2010) ressalta que pesquisadores coletam dados para diversos fins, usando vários métodos, podendo tanto a finalidade quanto os métodos influenciar no que consideram como "dados", e em que condições estes pesquisadores estão dispostos a compartilhar seus dados com os pares. Isso siginifica dizer que o conceito de dados de pesquisa pode variar não apenas de acordo com o domínio disciplinar, mas também de acordo com o propósito e até com a metodologia empregada na pesquisa. O presente trabalho foi uma tentativa de ajudar pesquisadores na classificação dos dados gerados por suas pesquisas, bem como na gestão e curadoria desses dados por bibliotecários e demais atores envolvidos no processo de gestão.

6 Referências

Berlin. (2003) *Declaration on open access to knowledge in the sciences and humanities*. *Berlin*, 2003. Disponível em: <http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf>. Acesso em: 20 dez. 2011.

Borgman, C. L (2010). Research data : who will share what, with whom, when an why. *Rratswd working paper* 161(10). Disponível em: <http://sydney.edu.au/research/data_policy/resources/ands_borgman_2010_research_data.pdf>. Acesso em: 19 maio 2013.

Buckland, Michael K (1991). Information as thing. *Journal of the american society for information science*, 42(5), 351-360.

Dahlberg, I (1978). A referent-oriented analytical concept theory of interconcept. *International classification, frankfurt*, 5(3), 142-150.

D-lib magazine. (2011). Disponível em: <http://www.dlib.org/dlib/january11/01contents.html> acesso em: 02 fev. 2019

D-lib magazine (2014). Disponível em: <http://www.dlib.org/dlib/january14/01editorial.html> acesso em: 02 fev. 2019

Harvey, Douglas Ross. *Digital curation:* a how-to-do-it manual. London:Facet, 2010.

Lyon, Liz (2007). Dealing with data; role, rigths, responsabilities and relationships. *Consultancy report*. 1-65. Disponível em: <http://opus.bath.ac.uk/412/1/dealing_with_data_report-final.pdf>. Acesso em: 19 maio 2013.

National Research Council – NRC (1999). *A question of balance: private rights and the public interest in scientific and technical databases*. Washington, dc: National Academy Press. 1999. Disponível em: <http://www.nap.edu>. Acesso em: 19 maio 2013.

National Science Foundation – NSF (2007). *Cyberinfrastructure vision for 21st century discovery.* Disponível em: <http://escience.caltech.edu/workshop/ci_vision_march07.pdf>. Acesso em: 19 maio 2013.

Organização para a cooperação e desenvolvimento econômico – OCDE (2007). *Principles and guidelines for access to research data from public data*. 2007. Disponível em: <http://www.oecd.org/dataoecd/9/61/38500813.pdf> acesso em: 17 fev. 2012.

Souza, Rosali Fernandez de (2012). Universo de ciência e tecnologia: organização e representação em classificações do conhecimento. In: Encontro Nacional de Pesquisa em Ciência da Informação - *ENANCIB, 13*. Rio de Janeiro, 2012. Disponível em: <http://www.eventosecongressos.com.br/metodo/enancib2012/arearestrita/pdfs/19371.pdf>. Acesso em: 20 mai 2013.

Wüster, E. (1981) L’étude scientifique qénérale de la terminologie, zone frontalière entre la linguistique, la logique, l’ontologie, l’informatique et les sciences des chose. In: rondeau, g.; felber, e. (org.). *Textes choisis de terminologie*. Québec: girserm, 1981, p. 57-114. (fondéments théoriques de la terminologie, v. I).