Análise do processo de recuperação de conjuntos de dados em repositórios governamentais

Analysis of datasets recovery process in government repositories

Fernando de Assis Rodrigues
Doutorando e Mestre em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP.
E-mail: fernando@elleth.org

Ricardo César Gonçalves Sant'Ana
Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP.
Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP.
ricardosantana@marilia.unesp.br

Edberto Ferneda
Doutor em Ciências da Comunicação pela Universidade de São Paulo – USP.
Professor do Departamento de Ciência da Informação da Universidade Estadual Paulista Julio Mesquita Filho – UNESP.
E-mail: ferneda@marilia.unesp.br

Resumo

O presente trabalho tem como objetivo identificar, na fase de recuperação, atributos disponíveis nos momentos em que se realiza pesquisas por conjuntos de dados em repositórios governamentais, a partir do modelo de Ciclo de Vida de Dados para a Ciência da Informação (CVD-CI) proposto por Sant'Ana (2013). A pesquisa fora delimitada a realização de buscas por conjuntos de dados através do mecanismo oferecido pelo sítio Portal Brasileiro de Dados Abertos, utilizando os termos 'Educação' e 'Saúde'. O uso do termo 'Saúde' resultou na recuperação de 14 conjunto de dados e o termo 'Educação' recuperou 23, totalizando 37 conjuntos de dados. A análise destes conjuntos de dados dividiu-se em duas etapas: na primeira foram identificados quais atributos estavam disponíveis na página contendo o resultado das buscas a partir termos utilizados. A segunda etapa consistiu em identificar os atributos disponíveis nas páginas referentes a cada um dos conjuntos de dados recuperados na busca. Como resultado, fora construído dois quadros: o primeiro identifica os atributos que estão disponíveis nas páginas com resultados da pesquisa pelo mecanismo de busca do site; o segundo, identifica os atributos disponíveis em cada conjunto de dados recuperado pela pesquisa. Os resultados demonstraram que na primeira etapa, não há diferença nos atributos disponíveis nos resultados de busca por ambos os termos. Entretanto, na segunda etapa houve discrepâncias nos atributos identificados em cada conjunto de dados.

Palavras-chave: Ciclo de Vida dos Dados. Coleta de Dados. Dados Abertos Governamentais. Repositório Governamental.

Abstract

The present study aims to identify, in the recovery stage, attributes available in moments when a user conducts datasets researches in government repositories, based on the Life Cycle Data Model for Information Science (CVD-CI) proposed by Sant'Ana (2013). The research was bounded out conducting searches for data sets offered through the search engine available on the site Brazilian Open Data Portal, using the terms 'education' and 'Health'. The use of the term 'health' resulted in the recovery of 14 datasets and the term 'education' recovered 23, totaling 37 datasets. Analysis of these datasets was divided into two stages: the first were identified which attributes were available on page containing the results of searches from terms used. The second step was to identify the attributes available on the pages for each datasets retrieved in the search. As a result, it was built two tables: the first identifies the attributes that are available on search results pages that were generated by site search engine. The second identifies the attributes available in each dataset retrieved by the search. The results showed that in the first stage, there is no difference in the attributes available in the search results by both terms. However, in the second stage there were discrepancies in the attributes identified in each dataset.

Keywords: Data Life Cycle. Data Gathering. Open Government Data. Governmental Repository.

InCID: R. Ci. Inf. e Doc., Ribeirão Preto, v. 6, n. 1, p. 38-56, mar./ago. 2015.

DOI: 10.11606/issn.2178-2075.v6i1p38-56

1. Introdução

A transparência das ações governamentais perante a sociedade é parte integrante nas discussões sobre tendências de modernização dos modelos de administração pública. Isso é reforçado, principalmente, no caso das democracias representativas, no qual os cidadãos elegem representantes diretamente ou indiretamente na composição dos poderes executivo e legislativo. (RODRIGUES; SANT'ANA, 2012b)

Esse novo modelo de administração pública

> […] busca redistribuir competências e recursos de coordenação entre diferentes níveis institucionais e organizacionais, governamentais e não-governamentais, permitindo o pluralismo institucional nas funções públicas, ao contrário do antigo modelo de monopólio estatal. (MALIN, 2006, p. 1)

A transparência das atividades e ações do Estado tem como uma de suas premissas fortalecer a participação dos cidadãos nesse novo modelo de administração pública. O fortalecimento pode ser garantido com a construção de ambientes democráticos que, dentre outras características, criem possibilidades de novos fluxos informacionais entre a administração do Estado e sociedade, garantindo assim uma maior visibilidade.

Nas democracias representativas, ampliar esses mecanismos de controle da sociedade civil sobre a administração pública, significa ir além do voto – ou seja – o comprometimento em criar condições para o acompanhamento social na administração pública além dos processos eleitorais. (BOHMAN, 1996)

A democracia brasileira regulamenta e autoriza o acesso dos dados governamentais pela sociedade.

> [...] todos têm direito a receber dos órgãos públicos informações de seu interesse particular, ou de interesse coletivo ou geral, que serão prestadas no prazo da lei, sob pena de responsabilidade, ressalvadas aquelas cujo sigilo seja imprescindível à segurança da sociedade e do Estado. (BRASIL, 1988, p. 1)

A lei conhecida como 'Lei de Acesso à Informação' (LAI) cria a obrigatoriedade do uso da infraestrutura da internet como instrumento de disseminação e acesso aos dados governamentais. A LAI estabelece que órgãos e entidades públicas terão de utilizar obrigatoriamente a internet como infraestrutura para a divulgação de dados e informações governamentais, via sítios oficiais do Estado. (BRASIL, 2011)

Em 2011, o governo brasileiro firmou parceria mediante uma iniciativa multilateral internacional de governo aberto (*Open Government Partnership - OGP*). O objetivo do *OGP*

(2011) é de unir esforços, em escala global visando garantir melhorias entre todos os parceiros. As melhorias propostas são baseadas na transparência, na melhoria de efetividade da administração pública e no aumento da responsabilidade dos governos em autorizar o acesso às informações governamentais pelos cidadãos.

Os parceiros da iniciativa responsabilizam-se por criarem metas para atingir esses objetivos – e periodicamente submeter o progresso das metas para a análise para um comitê independente. O progresso das metas exige a participação de lideranças políticas; investimentos em Tecnologias de Informação e Comunicação (TIC), bem como o conhecimento técnico dos artefatos; e a colaboração entre Estado e sociedade.

Dentre os diversos comprometimentos do cronograma assumido no plano de ação brasileiro, em parceria com o *OGP*, destacam-se:

a) A criação da Infraestrutura Nacional de Dados Abertos (INDA), que é um conjunto de tecnologias, processos, mecanismos de controle e padronização para o atendimento da legislação vigente do tema, bem como as conformidades estabelecidas nos padrões de interoperabilidade de governo eletrônico – o e-PING. (CONTROLADORIA-GERAL DA UNIÃO, 2006)

b) O Portal Brasileiro de Dados Abertos: sítio, implementado em dezembro de 2011, com o intuito de simplificar o acesso aos dados governamentais em âmbito federal.

Outro aspecto importante é que, devido a fatores tais como o barateamento de computadores, dispositivos de armazenamento e o próprio desenvolvimento contínuo das TIC, o volume de dados disponível por meio da infraestrutura da internet aumentou de forma muito expressiva. (RODRIGUES; SANT'ANA, 2012a)

Segundo Manyika et. al. (2011), estudos apontaram que no ano de 2010 o volume de novos dados gerados e armazenados por empresas e Estados foi de aproximadamente 7 *exabytes*. Somados os dados gerados pela sociedade este valor aumenta para 13 *exabytes*, ou seja, novos 13.958.643.712 *gigabytes* em dados no formato digital. Para 2020, a previsão é de um aumento de 44 vezes a quantidade de dados que fora armazenada digitalmente em 2009, com uma estimativa média da taxa de crescimento anual em 40%.

A definição de caminhos que contribuam para o acesso a esta quantidade crescente de dados disponíveis e ao atendimento a necessidades informacionais da sociedade é papel preponderante na Ciência da Informação. (RODRIGUES, 2012)

Entretanto, como o processo de recuperação de dados possui especificidades próprias, diferentes do processo de recuperação de informação via mecanismos de busca. Segundo Janowicz et. al. (2012), o estudo da recuperação de dados é o primeiro passo para os novos desafios e possibilidades que surgiram no processo de disponibilização de dados.

Para Van Rijsbergen (1999), a recuperação de dados difere-se da recuperação da informação em algumas propriedades, como na correspondência da pesquisa, na inferência, no modelo matemático, na classificação dos resultados, na linguagem utilizada para a elaboração de uma pesquisa, na recuperação dos dados e nas respostas aos possíveis erros no processo.

Para atender as especificidades do processo de recuperação com foco maior nos dados, este trabalho utiliza o modelo denominado Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI), proposto por Sant'Ana (2013).

> E para desempenhar esta missão, torna-se fundamental, conhecer e contribuir em todas as fases e fatores do processo de acesso a dados, o que leva a necessidade de se elaborar um modelo que sirva de base para compreensão sobre: quais são estas fases; como elas se relacionam; quais os fatores envolvidos em cada uma delas; quais os recursos disponíveis; como tirar o melhor proveito de cada uma delas, e; como é possível melhorá-las. (SANT'ANA, 2013, p. 2)

No CVD-CI, o processo do ciclo de vida dos dados está dividido em quatro fases (Figura 1): Coleta, Armazenamento, Recuperação e o Descarte. Cada uma destas fases são permeadas pelos fatores: preservação, disseminação, direitos autorais, qualidade, integração e privacidade.

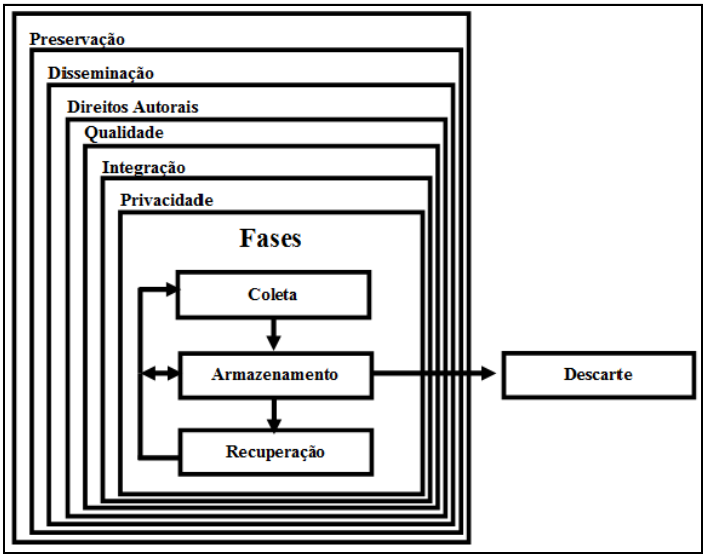


Figura 1 - Ciclo de Vida dos Dados para Ciência da Informação – (CVD-CI)
Fonte: Sant'Ana (2013)

A fase da coleta possui atividades vinculadas com o planejamento de obtenção, filtragem e organização de dados.

A fase de armazenamento está ligada aos processos e ações de persistência dos dados.

A fase de descarte está ligada a análise dos dados armazenados. Em algumas situações, esta fase pode gerar novos dados, via transferência dos dados descartados para novas bases, para efeito de preservação ou histórico.

A fase de recuperação é a fase que o acesso aos dados é concretizado, com atividades ligadas a consulta dos dados. Contudo, conjuntos de dados disponíveis na fase de recuperação podem ser passíveis de uma nova coleta para a geração de novos conjuntos de dados. Esta coleta pode ser de um agente interno ou externo a instituição que disponibiliza os dados.

Portanto, com dados cada vez mais presentes no cotidiano devido ao uso crescente de TIC e associados ao papel de destaque que o sítio Portal Brasileiro de Dados Abertos exerce no contexto de acesso à informação governamental; este trabalho tem o objetivo identificar na fase de recuperação, atributos disponíveis nos momentos em que se realiza pesquisas por conjuntos de dados no sítio Portal Brasileiro de Dados Abertos.

A pesquisa fora delimitada a realizar de pesquisas por conjuntos de dados através do mecanismo de busca oferecido pelo próprio sítio (Figura 2).

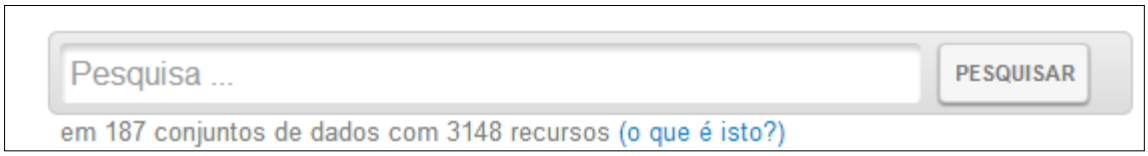


Figura 2 – Espaço destinado para pesquisas de conjuntos de dados
Fonte: BRASIL, 2004.

Na data da elaboração deste trabalho estavam disponíveis, no Portal Brasileiro de Dados Abertos, 187 conjuntos de dados, contendo 3148 recursos (Figura 2). Esta diferença de valores nos totais entre conjunto de dados e recursos deve-se que cada conjunto de dados pode possuir um ou mais recursos.

Como amostra para este trabalho foram utilizados somente conjuntos de dados recuperados pelo uso dos termos 'Saúde' e 'Educação' como expressões de busca. Foi escolhido o termo 'Saúde' pois a área de saúde pública brasileira está em evidência nos meios

de comunicação e no interesse público, principalmente em debates sob questionamentos da gestão financeira dos recursos, interligados com a qualidade do serviço oferecido. O termo 'Educação' também possui uma importância similar ao termo 'Saúde', gerando discussões e debates constantes sobre a educação pública brasileira na mídia, na sociedade e pelos representantes políticos, sob seus diversos aspectos. (RODRIGUES, 2012)

2. Metodologia

Como trata-se de uma análise da recuperação de dados, a metodologia adotada neste trabalho baseou-se no CVD-CI, especificamente nas questões de disseminação, inerentes à coleta de dados na fase de recuperação, identificando atributos nos conjuntos de dados disponíveis no sítio Portal Brasileiro de Dados Abertos.

Segundo Sant'Ana (2013), a existência de atributos que descrevam os dados disponíveis – e a sua preservação – são tão importantes quanto a própria preservação dos dados, pois são estes que permitem a interpretação do conteúdo ali disponível.

A análise exploratória dos conjuntos de dados disponíveis no sítio Portal Brasileiro de Dados Abertos, teve inicio a partir dos resultados obtidos pelos uso dos termos 'Educação' e 'Saúde' no mecanismo de busca. Esta análise dividiu-se em duas etapas.

Na primeira etapa, foram identificados quais atributos estavam disponíveis na página contendo o resultado das buscas a partir termos utilizados. Foram avaliados como atributos todas informações que consistiam-se como elementos descritivos de conjuntos de dados nos resultados disponíveis.

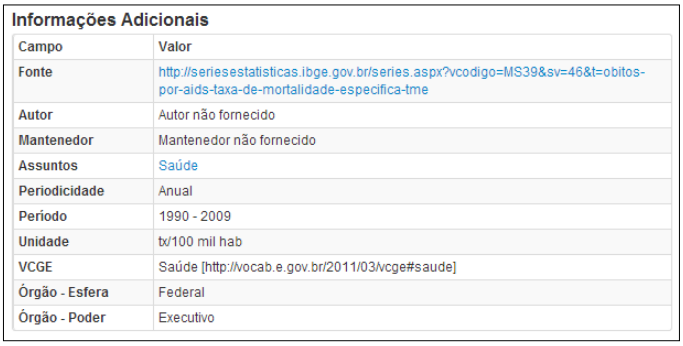
Para cada atributo, fora identificado o seu conteúdo (se o conteúdo deste elemento é um título, um resumo, uma data de atualização, um nome próprio de uma instituição, entre outros); o tipo de dado contido nesse atributo (se seu conteúdo é um texto, uma data, um *hiperlink*, um ícone, entre outros); e informações relacionadas ao conteúdo a partir da observação dos resultados.

A Figura 2 apresenta um recorte dos resultados obtidos através da busca do termo 'Educação' no Portal Brasileiro de Dados Abertos.

Figura 2 – Recorte dos resultados obtidos através da busca do termo 'Educação'
Fonte: BRASIL, 2004.

O resultado da busca permite o acesso à página de cada conjunto de dados por um *hiperlink*, que está rotulado pelo título do próprio conjunto de dados. Por exemplo, ao clicar no título 'Taxa de óbitos por AIDS', recuperado na busca da Figura 2, o sítio é direcionado para uma página contendo informações apenas do conjunto de dados 'Taxa de óbitos por AIDS'.

A segunda etapa consistiu em identificar os atributos que estão disponíveis nas páginas referentes a cada um dos conjuntos de dados recuperados na busca. Estas páginas possuem uma seção específica para estes atributos denominada 'Informações Adicionais', localizada na parte inferior. (Figura 3)



Informações Adicionais

Campo	Valor
Fonte	http://seriesestatisticas.ibge.gov.br/series.aspx?vcodigo=MS39&sv=46&t=obitos-por-aids-taxa-de-mortalidade-especifica-tme
Autor	Autor não fornecido
Mantenedor	Mantenedor não fornecido
Assuntos	Saúde
Periodicidade	Anual
Período	1990 - 2009
Unidade	tx/100 mil hab
VCGE	Saúde [http://vocab.e.gov.br/2011/03/vcge#saude]
Órgão - Esfera	Federal
Órgão - Poder	Executivo

Figura 3 – Seção 'Informações adicionais', encontrada nas páginas de cada conjunto de dados
Fonte: BRASIL, 2004.

Assim como na primeira etapa, para cada atributo fora identificado o seu conteúdo; o tipo de dado contido nesse atributo; e informações relacionadas ao conteúdo a partir da observação das páginas de cada conjunto de dados.

3. Resultados

O uso do termo 'Saúde' resultou na recuperação de 14 conjunto de dados e o termo 'Educação' recuperou 23, totalizando 37 conjuntos de dados.

O Quadro 1 apresenta os atributos identificados nos resultados das buscas pelos termos 'Educação' e 'Saúde'. As páginas com os resultados de ambas as pesquisas apresentaram os mesmos atributos.

Quadro 1 – Atributos identificados nos resultados de busca

Nome	Tipo de Dado
Título	Texto e *Hiperlink*
Descrição	Texto
Recursos	Texto e *Hiperlink*
Licença	Ícone ou Texto

Fonte: Autores

Nos resultados, foram identificados 4 atributos para descrever cada conjunto de dados recuperado pela pesquisa no sítio (Figura 4).

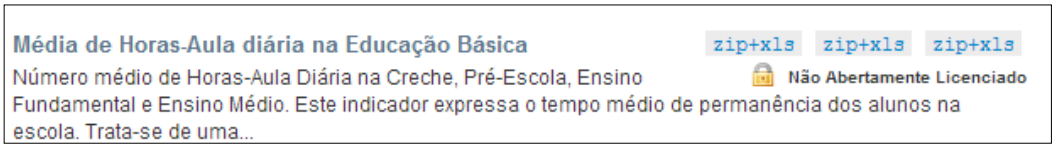


Figura 4 – Atributos que descrevem um conjunto de dados nos resultados de buscas
Fonte: Portal Brasileiro de Dados Abertos (2014)

O primeiro atributo é o título do conjunto de dados. O título é composto por um rótulo de um *hiperlink* que, ao ser acessado, redireciona o conteúdo para a página que contém com informações sobre o conjunto de dados escolhido.

Em seguida é exibido um campo com uma breve descrição sobre o conjunto de dados. Este campo é limitado para, no máximo, 28 palavras. Este texto é parte integrante das informações encontradas na página de cada conjunto de dados.

O terceiro atributo é composto pelos recursos contidos em cada conjunto de dados. Entretanto, para cada recurso em cada conjunto de dados é exibido apenas um *hiperlink* com o rótulo sendo o formato de arquivo do recurso. Por exemplo, se o recurso for um arquivo *Comma-Separeted Values* então o seu rótulo será 'csv'; se for um *Portable Document File*, o seu rótulo será 'pdf'.

O último atributo é relacionado ao tipo de licença atribuído ao uso de cada conjunto de dados. Este atributo é exibido como uma figura (ícone) de uma licença ou uma breve descrição.

Quando acessado a página com informações sobre um único conjunto de dados através do título nos resultados de busca, é encontrado no final da página uma seção denominada 'Informações Adicionais'. Esta seção contém exclusivamente atributos para auxiliar a contextualização tanto do próprio conjunto de dados em questão, como também de características dos recursos contidos no conjunto de dados.

O Quadro 2 apresenta a síntese dos atributos identificados na seção 'Informações Adicionais'.

Quadro 2 – Atributos identificados na seção 'Informações Adicionais' nos conjuntos de dados

Nome	Tipo de Dado	Informações sobre o conteúdo
Assuntos	*Hiperlink*	O atributo pode conter como conteúdo um ou mais assuntos. Cada um dos assuntos possui hiperlink. O rótulo representa a sua descrição e sua referência está vinculada ao Vocabulário Controlado do Governo Eletrônico (VCGE). Ao clicar no *hiperlink* é redirecionado para a página com todo o conteúdo do VCGE.
Atualidade	Data	O conteúdo da data de atualidade está no formato 'Mês/Ano' (Exemplo: 04/2013).
Autor	Texto ou *Hiperlink*	O conteúdo do atributo é apresentado na forma de *hiperlink* para o e-mail do autor. Caso o autor não seja informado, é apresentado uma mensagem textual com o conteúdo 'Autor não fornecido'.
cobertura espacial	Texto	O conteúdo, na forma textual, apresenta um nome do município, estado, região ou país.
Cobertura geográfica	Texto	O conteúdo, na forma textual, apresenta um nome do município, estado, região ou país.
cobertura temporal	Ano	Contém como conteúdo um ano, no formato de quatro dígitos.
Fonte	*Hiperlink*	*Hiperlink*, sendo o rótulo a *URL* do sítio que disponibilizou o conjunto de dados.
Granularidade	Texto	Conteúdo no formato texto. Na única ocorrência deste atributo o valor encontrado foi de "Instituição de ensino superior".
Granularidade geográfica	Texto	Conteúdo no formato texto, identificado por uma ou mais esferas do poder público (Federal, Estadual e Municipal) e/ou determinado tipo de entidade pública. Exemplo: "Escola".
Granularidade temporal	Texto	Conteúdo no formato de intervalos de tempo, tais como "Anual", "Bianual", "Semestral".
Mantenedor	Texto ou *Hiperlink*	O conteúdo do atributo é apresentado na forma de *hiperlink* para o e-mail do mantenedor. Caso o autor não seja informado, é apresentado uma mensagem textual com o conteúdo 'Mantenedor não fornecido'.
Órgão - Esfera	Texto	Conteúdo com uma ou mais esferas de poder (Federal, Estadual e/ou Municipal).
Órgão - Poder	Texto	Conteúdo com um ou mais poderes (Executivo, Legislativo e/ou Judiciário).
Periodicidade	Texto	Conteúdo no formato de intervalos de tempo, tais como "Anual", "Bianual", "Semestral".
Período	Data	Conteúdo, no formato de dois anos separados por um hífen, representando um intervalo de tempo. Exemplo: "1990 - 2009".

Nome	Tipo de Dado	Informações sobre o conteúdo
Unidade	Texto	O conteúdo representa uma unidade que determina uma escala utilizada nos dados contidos nos recursos do conjunto de dados. Exemplo: “Internações/100 000 hab”.
VCGE	Texto e *Hiperlink*	O atributo pode conter como valor um ou mais termos do VCGE. O atributo vincula o conjunto de dados com termos do VCGE. O atributo exibe sempre um termo do VCGE e, em seguida, a *URL* para o termo entre colchetes.

Fonte: Autores

Os atributos podem ser elementos temporais ('Atualidade', 'cobertura temporal', 'Granularidade temporal', 'Periodicidade' e 'Período') com a finalidade de relacionar uma unidade de tempo aos conjuntos de dados (quando foi elaborado, qual a última atualização, sobre qual o período trata-se os dados, etc.); elementos de autoridade ('Autor', 'Mantenedor', 'Órgão – Esfera', 'Órgão – Poder' e 'Fonte'), relacionando o conjunto de dados com entidades, órgãos, autores e esferas de poder); elementos de cunho geográfico ('cobertura espacial', 'Cobertura geográfica', 'Granularidade' e 'Granularidade geográfica') com o objetivo de identificar sobre qual região, município, estado ou país tratam-se aqueles dados; unidades de escala ('Unidade') mensurando qual escala foi adotada para apresentar os dados; e elementos interligando o conjunto de dados e assuntos preestabelecidos no Vocabulário Controlado do Governo Eletrônico ('VCGE' e 'Assuntos').

Cada atributo apresentado no Quadro 2 aparece em pelo menos uma página de um dos conjuntos de dados analisado. Porém, é importante ressaltar que não houve a ocorrência de um único conjunto de dados possuir todos os atributos identificados. Em suma, cada conjunto de dados contém um ou mais atributos do Quadro 2, porém não há nenhum conjunto de dados contendo todos os atributos.

O Quadro 3 exibe os atributos identificados na seção 'Informações Adicionais' em cada um dos 37 conjuntos de dados recuperados nas buscas. As linhas representam os conjuntos de dados e as colunas à direita são os atributos disponíveis. Quando um conjunto de dados contém determinado atributo, a célula de interseção está preenchida com o caractere 'X'. Na ausência do atributo, a célula de interseção não está preenchida.

Quadro 3 – Atributos identificados em cada conjunto de dados

Termo	Título do Conjunto de dados	Assuntos	Atualidade	Autor	cobertura espacial	Cobertura geográfica	cobertura temporal	Fonte	Granularidade	Granularidade geográfica	Granularidade temporal	Mantenedor	Órgão - Esfera	Órgão - Poder	Periodicidade	Período	Unidade	VCGE
Saúde	Municípios com Conselho Municipal de Saúde	X		X				X				X	X	X				
	Unidades Básicas de Saúde - UBS	X	X	X		X						X						X
	Aperfeiçoamento do Sistema Único de Saúde (SUS)	X		X				X				X	X	X				
	Taxa de incidência da dengue	X		X		X		X				X	X	X	X		X	X
	Doenças relacionadas ao saneamento ambiental inadequado - DRSAI	X		X		X		X				X	X	X	X	X	X	X
	Taxa de incidência de AIDS	X		X				X				X	X	X	X	X		X
	Taxa de óbitos por AIDS	X		X				X				X	X	X	X	X	X	X
	Estruturas da Fundacentro	X	X	X				X				X						X
	Postos de trabalho médicos por mil habitantes	X		X				X				X	X	X				
	Taxa de incidência de acidentes de trabalho em segurados da Previdência Social	X		X		X		X				X	X	X	X		X	X
	Equipamento de tomografia computadorizada por 100 mil de habitantes	X		X				X				X	X	X				
	Microdados do Registro Civil do estado de SP	X		X		X		X				X	X	X	X			X
	Tabela de Áreas de Conhecimento do Ensino Superior	X		X				X				X	X	X				X
	Índice Paulista de Responsabilidade Social - IPRS	X		X				X		X		X	X	X	X			X
Educa-ção	Municípios com Conselho Municipal de Educação	X		X				X				X	X	X				
	Microdados do Censo da	X		X				X				X	X	X	X			X

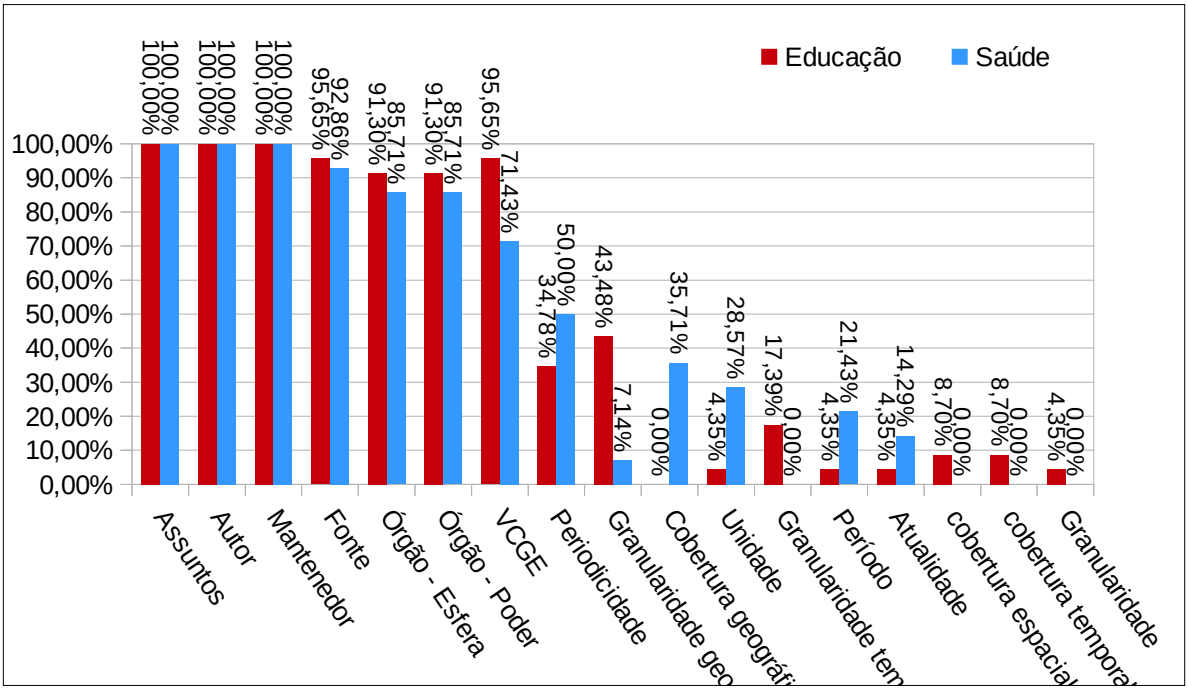
Termo	Título do Conjunto de dados	Assuntos	Atualidade	Autor	cobertura espacial	Cobertura geográfica	cobertura temporal	Fonte	Granularidade	Granularidade geográfica	Granularidade temporal	Mantenedor	Órgão - Esfera	Órgão - Poder	Periodicidade	Período	Unidade	VCGE
	Educação Superior																	
	Média de Alunos por Turma na Educação Básica	X		X				X		X		X	X	X				X
	Taxas de Rendimento Escolar na Educação Básica	X		X				X		X	X	X	X	X				
	Microdados da Pesquisa Nacional da Educação na Reforma Agrária - PNERA	X		X				X				X	X	X				X
	Microdados do Sistema Nacional de Avaliação da Educação Básica - Saeb	X		X				X				X	X	X	X			X
	Média de Horas-Aula diária na Educação Básica	X		X				X		X		X	X	X				X
	Taxas de distorção idade-série Escolar na Educação Básica	X		X						X		X	X	X				X
	Microdados do Censo Escolar	X		X				X		X	X	X	X	X				X
	Microdados Censo dos Profissionais do Magistério	X		X				X				X	X	X				X
	Taxa de Não Resposta no Censo Escolar	X		X				X		X		X	X	X	X			X
	Pesquisa de Controle de Qualidade do Censo Escolar 2011	X		X				X		X		X	X	X				X
	Instituições de Ensino Básico	X		X	X		X	X				X						X
	Microdados do Exame Nacional de Cursos (ENC-Provão)	X		X				X				X	X	X	X			X
	Instituições de Ensino Superior	X		X	X		X	X				X						X
	Microdados do Exame Nacional do Ensino Médio - Enem	X		X				X		X	X	X	X	X				X
	Taxa de analfabetismo	X	X	X				X				X	X	X				X

Termo	Título do Conjunto de dados	Assuntos	Atualidade	Autor	cobertura espacial	Cobertura geográfica	cobertura temporal	Fonte	Granularidade	Granularidade geográfica	Granularidade temporal	Mantenedor	Órgão - Esfera	Órgão - Poder	Periodicidade	Período	Unidade	VCGE
	funcional do Brasil de 2001 a 2009																	
	Microdados do Exame Nacional de Desempenho de Estudantes - Enade	X		X				X		X	X	X	X	X				X
	Microdados Prova Brasil	X		X				X				X	X	X	X			X
	Tabela de Áreas de Conhecimento do Ensino Superior	X		X				X				X	X	X				X
	Taxa de óbitos por AIDS	X		X				X				X	X	X	X	X	X	X
	Índice Paulista de Responsabilidade Social - IPRS	X		X				X				X	X	X	X			X
	Planilhas da Avaliação das Instituições de Nível Superior	X		X				X	X			X	X	X	X			X

Fonte: Autor

O Gráfico 1 exibe o percentual de identificação de cada um dos atributos, em relação ao total dos conjuntos de dados analisados, agrupando estes conjuntos de dados pelos termos utilizados em sua recuperação ('Saúde' e 'Educação'). Para calcular o percentual, dividiu-se a quantidade de vezes que o atributo foi identificado nos conjuntos de dados de um termo pelo total de conjuntos de dados recuperados em cada termo. Utilizou-se um arredondamento de duas casas decimais no valor do percentual.

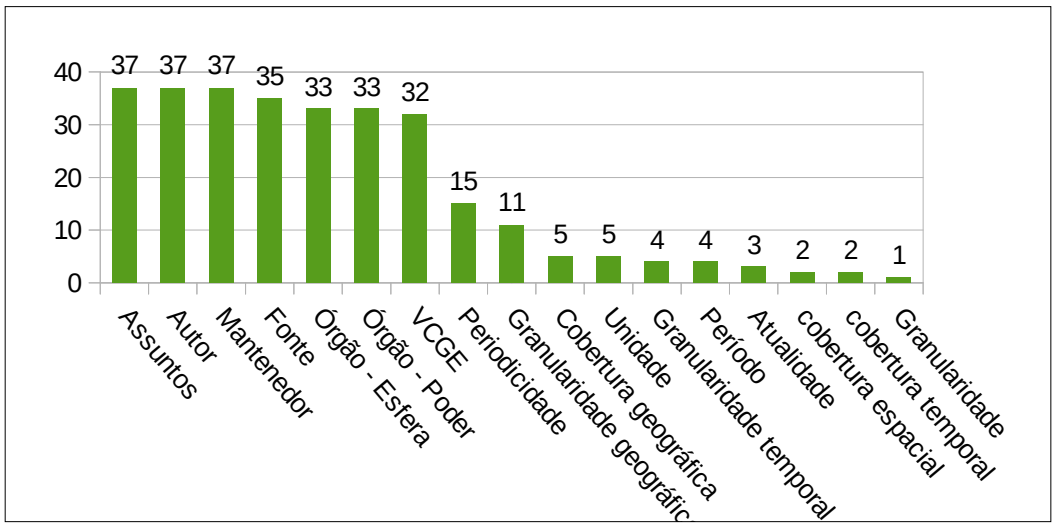
Gráfico 1 – Percentual de identificação de atributos, agrupados por termos



Fonte: Autores

Por exemplo, o atributo 'Unidade' foi identificado em 4 conjuntos de dados recuperados pelo termo 'Saúde'. Esta quantidade (4) representa que em 28,57% dos conjuntos de dados recuperados pelo termo 'Saúde' está disponível o atributo 'Unidade'.

Gráfico 2 – Quantidade de atributos identificada nos conjuntos de dados da pesquisa



Fonte: Autores

O Gráfico 2 exibe a quantidade de vezes que os atributos foram identificados em cada um dos conjunto de dados. Cada coluna representa um único atributo e quão mais alta a coluna, mais o atributo aparece nos conjuntos de dados selecionados para o universo desta pesquisa. Quando um atributo é identificado em 37 conjuntos de dados, significa que este é

identificado em todos conjuntos de dados participantes no universo desta pesquisa.

4. Conclusões

Na primeira etapa, que contém os resultados de busca por um termo de pesquisa, não há discrepâncias entre os atributos identificados em cada conjunto de dados. Ou seja, independente do termo utilizado na pesquisa através do mecanismo de busca, os atributos 'Título', 'Descrição', 'Recursos' e 'Licença' estão disponíveis em todos os conjuntos de dados recuperados.

Contudo, no atributo 'Recursos' é difícil ao usuário o entendimento de o que é cada um dos recursos disponíveis em um conjunto de dados, pois este atributo apresenta apenas como descrição dos recursos disponíveis o formato do arquivo como rótulo de *hiperlink* para o acesso direto ao mesmo. A Figura 5 apresenta um exemplo dos rótulos para o acesso direto aos recursos, destacados por um retângulo de borda vermelha.

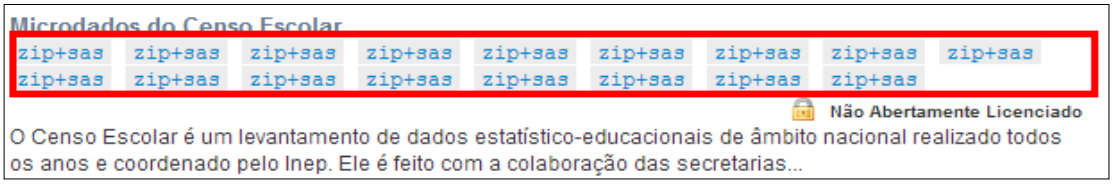


Figura 5 – Rótulos de acesso direto à recursos de um conjunto de dados
Fonte: Autores

Outro aspecto importante é a falta de padronização entre os tipos de licença. Em alguns conjuntos de dados, a licença é exibida em um logotipo em formato de ícone e em outros, por um cadeado seguido de um texto. A Figura 6 exibe as diferenças entre a exibição de licenças nos resultados de busca, apontando-as com setas vermelhas.

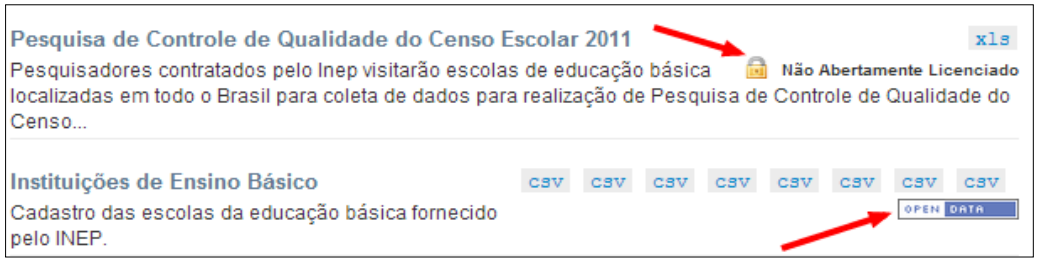


Figura 6 – Diferenças entre exibições de licenças nos resultados de busca
Fonte: Autores

Os resultados da busca poderiam incluir outros atributos já existentes nos conjuntos de dados. Estes novos atributos auxiliariam o processo de busca, o que facilitaria na identificação da origem dos dados, já que existem conjuntos de dados que aparecem em pesquisas com ambos os termos ('Educação' ou 'Saúde') como o conjunto 'Taxa de óbitos por AIDS'.

Há atributos que aparecem preenchidos na maioria dos conjuntos de dados identificados, e que poderiam estar disponíveis nos resultados de busca. Por exemplo, 35 dos 37 conjuntos de dados já contém o atributo 'Fonte' (Gráfico 2).

Na segunda etapa, onde são analisados as páginas com informações de cada conjunto de dados, há maiores discrepâncias. Com exceção dos atributos 'Assuntos', 'Autor' e 'Mantenedor', não há nenhum outro atributo que esteja disponível em todos os conjuntos de dados. Além disso, apesar do atributo 'Assuntos' estar disponível em alguns conjuntos de dados, o atributo está em branco pois não foi ainda criado o vínculo entre estes conjuntos de dados e os termos disponíveis no VCGE.

Não há similaridade entre os atributos disponíveis nas páginas dos conjuntos de dados quando estes estão agrupados pelos termos pesquisados para a sua recuperação. Isto representa que conjuntos de dados recuperados pelo termo 'Educação' podem apresentar uma coleção de atributos diferentes dos conjuntos de dados recuperados pelo termo 'Saúde'; e vice-versa.

Por exemplo, os atributos 'VCGE', 'Granularidade geográfica', 'Cobertura geográfica' e 'Unidade', identificados nos conjuntos de dados e agrupados pela sua recuperação (recuperados pelo termo 'Educação' e pelo termo 'Saúde') diferem entre si em um percentual maior de 24%. Com base nos dados coletados, entende-se que essa diferença pode ocorrer devido a falta de obrigatoriedade de preenchimento dos atributos no momento de incluir novos conjuntos de dados no sítio, já que os dados encontrados no sítio são oriundos de diversas fontes, órgãos e áreas do Estado brasileiro.

Os atributos 'cobertura temporal' e 'cobertura espacial' não estão padronizados com os demais, aparecendo em todos os momentos com as primeiras letras minúsculas.

O conteúdo do atributo 'VCGE' repete-se no atributo 'Assuntos', porém não está claro o que é seu significado. As informações sobre o significado dos termos do VCGE só são encontradas quando copia-se o conteúdo do atributo em uma barra de endereços de um navegador de internet. Este procedimento manual faz com que o conteúdo do atributo, que é

uma URL, crie o acesso ao sítio oficial do VCGE.

Não estão claros nos atributos 'Órgão – Esfera' e 'Órgão – Poder' se o seu conteúdo representa que poderes e esferas que vinculadas com o responsável por disponibilizar o conjunto de dados; ou se seu conteúdo representa poderes e esferas que os dados ali contidos terão vínculo.

O trabalho contribui na explicitação de ocorrências de diferentes conjuntos de metadados disponíveis nas etapas do processo de recuperação de dados contidos Portal Brasileiro de Dados Abertos. Como sugestão de trabalhos futuros, esta análise poderá ser aplicada em outros cenários e contextos de dados públicos, tais como sítios de pertencentes a outras nacionalidades, e sítios brasileiros de outras esferas e poderes.

Referências

BOHMAN, J. **Public deliberation, pluralism, complexity and democracy.** London: MIT Press, 1996.

BRASIL. **Constituição da República Federativa do Brasil de 1988.** Portal do Planalto, Brasília, DF. Disponível em: <www.planalto.gov.br/ccivil_03/Constituicao/Constituicao.htm>. Acesso em: 05 jan. 2014.

_____. Lei número 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. **Portal do Planalto**, Brasília, DF, 18 nov. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm>. Acesso em: 05 jan. 2014.

_____. **Portal brasileiro de dados abertos.** Brasília, 2014. Disponível em: <http://dados.gov.br>. Acesso em: 04 jan. 2014.

CONTROLADORIA GERAL DA UNIÃO. **Plano de ação do governo brasileiro:** parceria para governo aberto (OGP). Brasília, 2006. Disponível em: <http://www.cgu.gov.br/acessoainformacao/destaques/ogp/ogp-brazil-actionplan.pdf>. Acesso em: 04 jan. 2014.

JANOWICZ, K. et al. Geospatial semantics and linked spatiotemporal data: past, present, and future**. Semantic Web**, v. 3, n. 4, p. 321-332, 2012.

MALIN, A. M. B. Gestão da informação governamental: em direção a uma metodologia de avaliação. **DataGramaZero**, v. 7, n. 5, out. 2006. Disponível em: <http://www.dgz.org.br/out06/Art_02.htm>. Acesso em: 04 jan. 2014.

MANYIKA, J. et. al. **Big data:** the next frontier for innovation, competition and productivity**.** Nova Iorque: McKinnsey Global Institute, 2011. 156 p.

OPEN GOVERNMENT PARTNERSHIP. **Open goverment partnership web site**. 2011. Disponível em: <http://www.opengovpartnership.org>. Acesso em: 05 jan. 2014.

RODRIGUES, F. A. **Mapeamento de tecnologias informacionais sobre dados abertos em saúde pública**: destino de repasses financeiros federais. 2012. 143 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Marília, 2012.

_____.; SANT'ANA, R. C. G. Restrições tecnológicas e de acesso a dados disponíveis sobre destinos de repasses financeiros federais para a saúde pública em ambientes informacionais digitais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 13., 2012. **Anais digitais.** Rio de Janeiro: FIOCRUZ, 2012a. Disponível em: <http://www.eventosecongressos.com.br/metodo/enancib2012/arearestrita/pdfs/19435.pdf>. Acesso em: 10 jan. 2014.

_____.; _____. G. Uso de modelos de dados multidimensionais para ampliação da transparência ativa**. LIINC em Revista**, v. 9, n. 2, nov., 2012b. Disponível em: <http://revista.ibict.br/liinc/index.php/liinc/article/viewFile/599/428>. Acesso em: 05 nov. 2013.

SANT'ANA, R. C. G. Ciclo de vida dos dados e o papel da ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013**. Apresentações.** Florianópolis: UFSC, 2013. ISBN 978-85-65044-06-6. Disponível em: <http://enancib2013.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319>. Acesso em: 29 jan. 2014.

VAN RIJSBERGEN, C. J. **Information retrieval.** 2. ed. Londres: Butterworths, 1999. Disponível em: <http://www.dcs.gla.ac.uk/Keith/Preface.html>. Acesso em: 30 ago. 2014.

Artigo submetido em: 02 fev. 2014
Artigo aceito: 02 out. 2014