Descrição de conjuntos de dados na *Web* com *Schema.org*

Marcos Teruo Ouchi
Universidade Federal de São Carlos – UFSCar, e-mail: ouchi@isci.com.br

Ana Carolina Simionato
Universidade Federal de São Carlos – UFSCar, e-mail: acsimionato@ufscar.br

Resumo
O reuso dos dados contribui para minimizar a duplicação do trabalho de coleta, otimizando custos e recursos; possibilitando a preservação de longo prazo e mantendo a integridade dos dados; também fornece salvaguardas contra má conduta científica, incluindo fraudes e ferramentas de treinamentos para novas gerações de pesquisadores. Dessa forma, esse trabalho contextualiza-se pelo questionamento de como descrever os conjuntos de dados produzidos nas pesquisas científicas de forma a permitir a descoberta e recuperação desses conjuntos de dados de pesquisa na *Web*? Para isso, objetiva-se a analisar a iniciativa de representação de recursos digitais para enriquecimento de informações *Schema.org* como alternativa para representação dos dados. De natureza exploratória, abordagem qualitativa, apresenta um levantamento bibliográfico e análise de exemplos. Por se tratar de uma iniciativa recém lançada e em fase de desenvolvimento, mas com grande potencial de aplicação com resultados práticos, o padrão *Schema.org* para dados de pesquisa enfrenta inúmeros desafios apresentando-se como uma proposta inovadora e potencialmente funcional, podendo, quando totalmente implementado, tornar acessíveis os conjuntos de dados de pesquisa na *Web*.

Palavras-chave: Enriquecimento de dados. *Schema.org*. Gestão de dados científicos.

INTRODUÇÃO

A publicação científica vem se consolidando durantes os séculos, culminando com o atual estágio da comunicação científica por meio das instituições e agências financiadoras de pesquisas, de eventos, das editoras, periódicos e repositórios, que se vêm atualmente frente aos desafios decorrentes da evolução, da redução dos custos e aumento exponencial de produção de objetos digitais.

Nesse contexto, os recentes avanços tecnológicos transformaram a maneira de como publicar, abarcando a forma textual e todos os objetos relacionados, denominando assim, a publicação ampliada. Esse tipo de publicação permite que os conjuntos completos de dados sejam disponibilizados conjuntamente às publicações em repositórios de dados digitais disponíveis na *internet* e, por universidades e editoras.

A digitalização, a utilização de equipamentos digitais de coletas de dados, a virtualização, uso de dispositivos embarcados, *internet* das coisas (iOT) entre outras, abriram

amplas possibilidades de coleta de dados de comportamentos sociais, antes, de uso extremamente custoso, trabalhoso e que demandava muitos recursos humanos, financeiros e de tempo para sua coleta, manutenção, utilização e reutilização. Da mesma forma, os dados produzidos em pesquisas apresentam-se novamente como uma preocupação para o progresso científico, uma vez que os dados são e sempre foram essenciais para estabelecer a validade, reprodução e replicação da pesquisa.

Neste sentido, faz-se necessário o desenvolvimento de ferramentas, métodos e medidas apropriadas de preservação para que dados oriundos de pesquisas científicas não se percam e sejam recuperados, evitando-se o que ocorre quando armazenados em suportes como mídias magnéticas obsoletas.

O reuso dos dados contribui para minimizar a duplicação do trabalho de coleta, otimizando custos e recursos; possibilitando a preservação de longo prazo mantendo a integridade dos dados; fornece salvaguardas contra má conduta científica, incluindo fraudes e ferramentas de treinamentos para novas gerações de pesquisadores.

A partir da temática abordada, questiona-se como descrever os conjuntos de dados produzidos nas pesquisas científicas para que possibilite a descoberta e recuperação na *Web*? Como objetivo, busca-se analisar a iniciativa de representação de recursos digitais para enriquecimento de informações *Schema.org*.

O motivo da escolha deste padrão provém do fato de que, hoje, 99,8% das buscas realizadas na *Web* são feitas através dos buscadores envolvidos na criação do vocabulário *Schema.org* (STATISTA, 2017). Assim, como pesquisa qualitativa, teórica, prática, exploratória e pautada em Gil (2007), foi realizado um levantamento bibliográfico e uma análise do *Schema.org* com exemplos para a descrição de conjuntos de dados na *Web*.

2 DADOS CIENTÍFICOS E *SCHEMA.ORG*

Os dados científicos apresentam-se novamente como preocupação para o progresso científico, em razão de que os dados são e sempre foram essenciais para estabelecer a validade, reprodução e replicação da pesquisa. A digitalização de documentos, a utilização de equipamentos de coletas de dados, a virtualização, uso de dispositivos embarcados, abriram amplas possibilidades de coleta de dados de comportamentos sociais, até pouco tempo, de uso extremamente custoso e trabalhoso e que demandava muitos recursos humanos, financeiros e de tempo para sua manutenção e utilização.

Hoje, "Vivenciamos o quarto paradigma da ciência, o qual tem redefinido o *modus operandi* das práxis científicas como consequência dos desafios impostos pela produção de dados em larga escala". (CURTY; CERVANTES, 2016, p. 1).

> O desenvolvimento de novos aparatos científicos, instrumentos, sensores, escalas e o uso intensivo de modelos de simulação da natureza geram uma quantidade imensa de dados, delineiam as fronteiras de um novo paradigma científico – conhecida como *eScience* -, criam novas metodologias de produção de conhecimento científico, formas de compartilhamento e de socialização entre os pesquisadores e alteram o fluxo tradicional de comunicação científica e de revisão por pares. (SAYÃO; SALES, 2017, p. 67).

"Dados provêm evidências para o conhecimento científico publicado, que é a fundação para todo o progresso científico" (MOLLOY, 2011, p. 1), sendo essenciais, não apenas para o desenvolvimento, mas também para estabelecer a validade, reprodução e replicação da pesquisa. Neste sentido, tornando-se um valioso produto da pesquisa. Mecanismos como a revisão por pares e pesquisas de reprodutibilidade podem atribuir validade aos resultados apresentados em formatos tradicionais de divulgação científica, entretanto, as dificuldades para obtenção dos conjuntos de dados utilizados são notórias, trazendo à tona uma evidente preocupação com o estado de preservação desses conjuntos de dados. Harvey (2010, p. 10) afirma:

> As respostas a essas ameaças de continuidade digital que são baseadas em abordagens de preservação tradicionais não funcionam. Simplesmente capturar dados em mídias de armazenamento estáveis e depois copiá-las em novas mídias quando as anteriores se tornarem obsoletas é uma ameaça. Dados digitais devem ser geridos desde sua criação se a pretensão for de sobrevivência.

Os avanços tecnológicos permitem que, atualmente, os conjuntos completos de dados sejam disponibilizados facilmente e com baixos custos, fazendo-se necessário o desenvolvimento de ferramentas, metodologias e medidas apropriadas de preservação para que dados oriundos de pesquisas científicas não se percam, como já ocorrem quando armazenados em suportes como mídias magnéticas obsoletas.

O ciclo de vida de curadoria dos dados são processos contínuos que fornecem uma visão geral e de alto nível dos estágios envolvidos no gerenciamento e preservação de dados para uso e reutilização. Segundo a *Digital Curation Centre* (2013) os ciclos assumem as seguintes etapas: etapa conceitual: criação, acesso e uso, avaliação e seleção, descarte, ingestão,

preservação, revalorização, armazenamento, reutilização e transformação. Em 2014, a FORCE11 publicou um manifesto assinado por grandes editores onde, além de outros compromissos, assume que "[...] os dados devem ser considerados produtos de pesquisa legítimos e passíveis de citação" (DATA CITATION SYNTHESIS GROUP, 2013).

Não surpreende tal posicionamento, uma vez que o trabalho de coleta, higienização, organização, estruturação e publicação de conjuntos de dados ou *datasets* munidos de suas respectivas descrições (metadados) pode ser considerado produto legítimo de esforço de um ou mais pesquisadores e passível de reconhecimento pela comunidade acadêmica como um produto de pesquisa.

Nas abordagens para a gestão do ciclo de vida dos dados de pesquisa, a documentação por meio de metadados é tratada como fator preponderante para fins de acesso, compreensão, compartilhamento e preservação dos conjuntos de dados, pois

> Os metadados explicitam os diferentes aspectos do recurso que descreve: sua estrutura, conteúdo, qualidade, contexto, origem, propriedade e condição. E auxiliam na organização, favorecem a interatividade, validam as identificações e asseguram a preservação e principalmente, otimizam o fluxo informacional melhorando o acesso aos dados e a localização dos recursos informacionais. (SANTOS; SIMIONATO; ARAKAKI, 2014, p. 150).

Portanto, "A qualidade e precisão dos esquemas de metadados adotados e o rigor da sua aplicação são de crucial importância na garantia de que as coleções de dados possam ser acessadas, usadas e reutilizadas interdisciplinarmente pelo tempo que for necessário". (SAYÃO; SALES, 2013, p. 17).

Ainda, segundo Sayão e Sales (2015, p. 28), "[...] as exigências sobre o nível de descrição e de atribuição de metadados devem ser identificadas desde o começo do seu projeto e revistas ao longo do ciclo de vida dos seus dados, sendo essa a essência de uma boa curadoria de dados".

> A preservação segura vai garantir a conservação da informação ao longo do tempo, mas para isso devem-se usar as melhores práticas em todo o ciclo de vida dos dados: criação, descrição e curadoria, além de seu armazenamento seguro, que garanta seu uso e reuso. (SALES; CAVALCANTI, 2015, p. 92).

Portanto, se o objetivo do pesquisador for o de disponibilizar seus conjuntos de dados na *World Wide Web*, para recuperação a partir dos principais mecanismos de busca de informações na *Web*, será necessário considerar a descrição desses *datasets* escolhendo, desde

o início, um ou mais padrões para vocabulários que seguem padrões internacionais de metadados, como o *Schema.org*.

A partir de uma iniciativa de buscadores em junho de 2011, o *Google*, o *Yahoo*, *Bing* e *Yandex* idealizaram o *Schema.org* como uma proposta compartilhada e colaborativa com o objetivo de criar, manter e promover esquemas de dados estruturados para Internet. (SCHEMA.ORG, 2011).

A estrutura do *Schema.org* busca melhorar a busca de informações e estima-se que dezenas de milhares ou até mesmo milhões de *datasets* estejam disponíveis em muitas centenas de repositórios de dados na *Web* (NOY, 2017) tornando uma tarefa desafiadora a localização, garantia de proveniência, veracidade e identificação desses conjuntos de dados.

Diferentemente das outras iniciativas na qual buscou inspiração, como *Linked Open Data*, *cyc*, entre outros, o *Schema.org* não prevê a construção de ontologias ou bases de conhecimentos, mas uma maneira dos buscadores envolvidos em seu desenvolvimento proverem uma experiência mais rica nos resultados das buscas para seus usuários.

Com o objetivo de promover um ecossistema de publicação, consumo e descoberta de conjunto de dados, a iniciativa *Schema.org* disponibilizou recentemente, por solicitação da própria comunidade de usuários, um conjunto de metadados denominado *datasets* como parte do conjunto *creative works*, partindo da premissa de que um conjunto de dados pode ser representado da mesma forma que qualquer outro conceito como livros, filmes, receitas e pessoas.

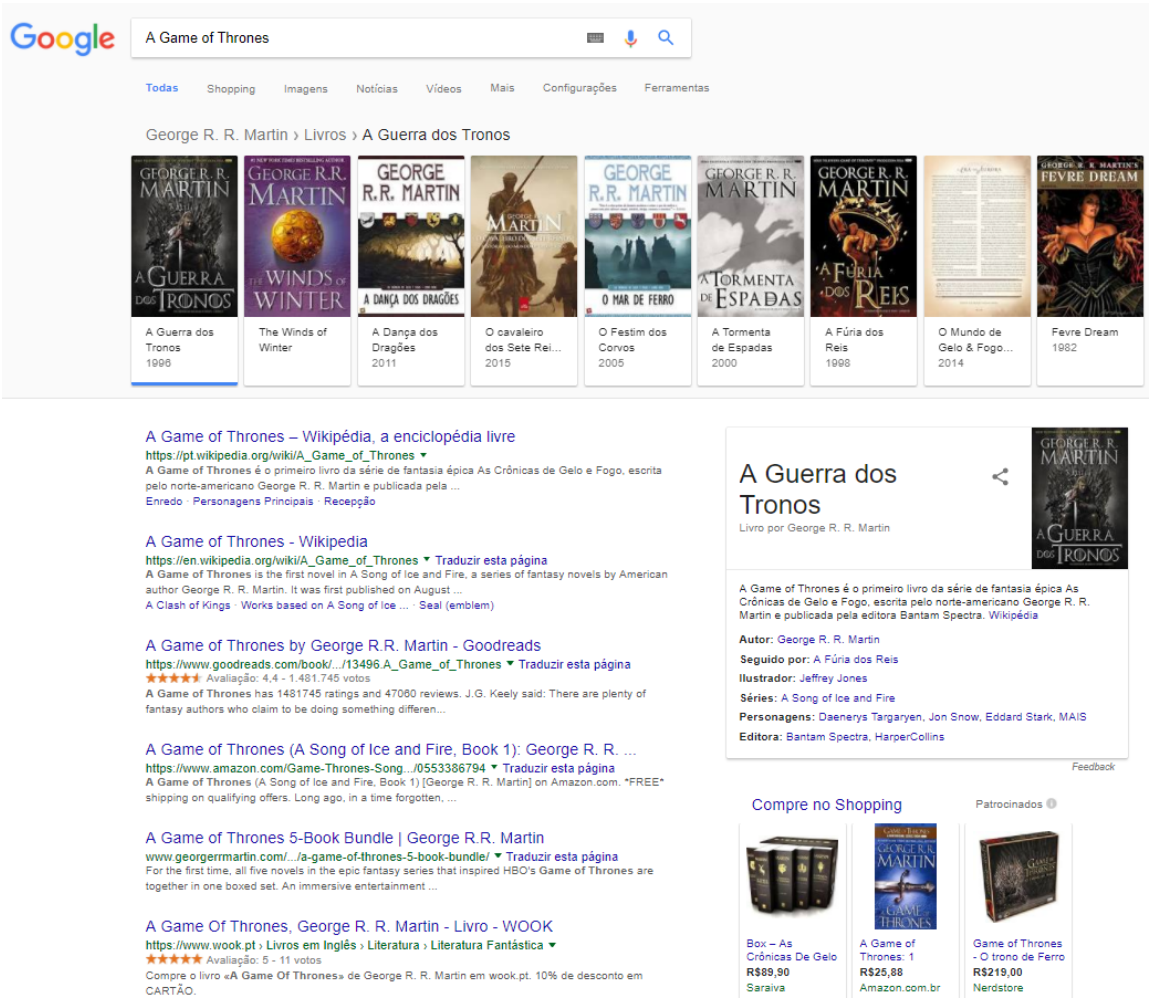
A Google (2017), uma das mantenedoras do projeto *Schema.org*, qualifica de forma ampla um conjunto de dados como:

- Uma tabela ou um arquivo CSV com alguns dados.
- Um arquivo em um formato proprietário que contenha dados.
- Uma coleção de arquivos que juntos constituem um conjunto de dados significativos.
- Um objeto estruturado com dados em algum outro formato que se possa processar em uma ferramenta de processamento.
- Imagens que contenham dados.
- Qualquer outra coisa que se pareça com dados.

A marcação de dados e de conjuntos de dados permite o reconhecimento desses conjuntos de dados atribuindo-lhes valor semântico, reduzindo "[...] a ambiguidade sobre o que as páginas descrevem tornando a integração dos dados nos motores de busca mais eficiente".

(FONS; PENKA; WALLIS, 2012) permitindo que os resultados de busca sejam relacionados e visualizados de forma mais adequada, de acordo com o tipo de dados que estejam sendo apresentados em seus resultados de busca como pode ser observado na figura 1, onde, a título de exemplo, o termo "*A Game of Thrones*" resultou em uma estrutura visual onde cada manifestação da obra foi apresentada de forma diferenciada.

Figura 1 - Resultado de pesquisa com *Rich Snippets*



Fonte: Google (2017).

A *Google* denomina o formato de apresentação enriquecido de informações como *Rich Snippets*, que modifica a maneira como apresenta um conjunto de informações classificando-o de acordo com sua natureza.

A título de exemplo, na figura 2, os dados de diversas fontes foram utilizados para condensar, em um quadro visualmente mais atrativo do que simples texto em formato de lista,

informações sobre o autor Ziraldo Alves Pinto. Na figura 2 foram apresentadas imagens, dados biográficos do autor, relação de obras mais relevantes para quem realizou a pesquisa e uma relação de outros autores "enriquecendo" desta maneira a experiência de busca do pesquisador.

Figura 2 - *Rich Snippet* de autor

Fonte: Google (2017).

Cada conjunto de propriedades correspondem a uma classe do *Schema.org* que pode ser representada visualmente de forma diferenciada nos resultados dos buscadores e aplicativos que utilizam o vocabulário, como no exemplo da figura 3, onde os metadados atribuídos correspondem ao subconjunto de "*Thing* > *CreativeWork* > *Article* > *NewsArticle*" permitindo ao buscador apresentar as informações num formato diferenciado de outros entes.

Figura 3 – Exemplo do *Rich Snippets NewsArticle*

Fonte: Google (2017).

Até a data de elaboração deste trabalho, nenhum dos principais buscadores havia implementado um formato de apresentação diferenciado para os *datasets* de trabalhos oriundos de pesquisa científica, entretanto, iniciativas como o *data.gov*, um catálogo de dados abertos do governo dos Estados Unidos da América, disponibiliza conjuntos de dados descritos com o padrão *Schema.org* conforme apresentado na figura 4.

O código apresentado descreve o conjunto de dados de eventos de tempestades contendo estatísticas sobre danos pessoais e estimativas de danos ocorridos nos Estados Unidos da América entre 1950 até os dias atuais (NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION, DEPARTMENT OF COMMERCE, 2017). O conjunto de dados contêm uma listagem cronológica, por estado, de furacões, tornados, tempestades, granizo, inundações, condições de seca, raios, ventos fortes, neve, temperaturas extremas e outros fenômenos climáticos.

O exemplo foi escrito em *ld-json* um dos formatos disponíveis para ligação de dados (dados vinculados), sendo um formato leve, facilmente entendível por seres humanos e máquinas. Baseia-se no formato *json* e é ideal para a interoperabilidade de dados.

Figura 4 - Conjunto de dados representado utilizando *Schema.org* em LD e JSON

dataset_example.json

```
<script type="application/ld+json">
{
  "@context":"http://schema.org/",
  "@type":"Dataset",
  "name":"NCDC Storm Events Database",
  "description":"Storm Data is provided by the National Weather Service (NWS) and contain statistics on...",
  "url":"https://catalog.data.gov/dataset/ncdc-storm-events-database",
  "sameAs":"https://gis.ncdc.noaa.gov/geoportal/catalog/search/resource/details.page?id=gov.noaa.ncdc:C00510",
  "keywords":[
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > CYCLONES",
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > DROUGHT",
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > FOG",
     "ATMOSPHERE > ATMOSPHERIC PHENOMENA > FREEZE"
  ],
  "creator":{
     "@type":"Organization",
     "url": "https://www.ncei.noaa.gov/",
     "name":"OC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce",
     "contactPoint":{
        "@type":"ContactPoint",
        "contactType": "customer service",
        "telephone":"+1-828-271-4800",
        "email":"ncei.orders@noaa.gov"
     }
  },
  "includedInDataCatalog":{
     "@type":"DataCatalog",
     "name":"data.gov"
  },
  "distribution":[
     {
        "@type":"DataDownload",
        "encodingFormat":"CSV",
        "contentUrl":"http://www.ncdc.noaa.gov/stormevents/ftp.jsp"
     },
     {
        "@type":"DataDownload",
        "encodingFormat":"XML",
        "contentUrl":"http://gis.ncdc.noaa.gov/all-records/catalog/search/resource/details.page?id=gov.noaa.ncdc:C00510"
     }
  ],
  "temporalCoverage":"1950-01-01/2013-12-18",
  "spatialCoverage":{
     "@type":"Place",
     "geo":{
        "@type":"GeoShape",
        "box":"18.0 -65.0 72.0 172.0"
     }
  }
}
</script>
```

Fonte: National Oceanic and Atmospheric Administration, Department Of Commerce (2017)

O mesmo conjunto de dados também pode ser descrito em outros formatos como *microdata* ou *RDFa*, tornando-se, assim, muito flexível para diversos usos em diferentes linguagens de programação e de intercâmbio de dados.

As propriedades básicas definidas para o conjunto de dados pelo *Schema.org*, dos quais o uso do maior número deles é aconselhado desde que pertinente, como apresentado no quadro 1.

Quadro 1 – Propriedades básicas definidas para o conjunto de dados

Propriedade	Tipo	Descrição
name	Texto	Um nome descritivo de um conjunto de dados (por exemplo, "profundidade de neve no hemisfério norte")

description	Texto	Um breve resumo descrevendo um conjunto de dados.
url	URL	Localização de uma página que descreve o conjunto de dados.
sameAs	URL	Outros URLs que podem ser usados para acessar a página do conjunto de dados.
version	Texto, Número	O número da versão para este conjunto de dados.
keywords	Texto	Palavras-chave que resumem o conjunto de dados.
variableMeasured	Texto, Valor da Propriedade	O que o conjunto de dados mede? (por exemplo, temperatura, pressão)

Fonte: Google (2017)

Como exposto no quadro 2, é possível descrever informações adicionais sobre a publicação do conjunto de dados, como a licença, quando foi publicado, seu identificador de objeto digital (DOI) ou apontar para uma versão canônica do conjunto de dados em um outro repositório.

Quadro 2 – Informações de proveniência ou licença

Propriedade	**Tipo**	**Descrição**
license	URL, Texto	Uma licença sob a qual o conjunto de dados é distribuído.
identifier	URL, Texto, Valor da Propriedade	Um identificador para o conjunto de dados, como um DOI.
sameAs	URL	Um link para uma página que fornece mais informações sobre o mesmo conjunto de dados, geralmente em um repositório diferente.

Fonte: Google (2017)

Os conjuntos de dados são frequentemente publicados em repositórios que contêm muitos outros conjuntos de dados. O mesmo conjunto de dados pode ser incluído em mais de um repositório, por meio do conjunto de propriedades do catálogo de dados é possível se referir diretamente a ele, destacado pelo quadro 3.

Quadro 3 - Propriedade do catálogo de dados

Propriedade	**Tipo**	**Descrição**
includedInDataCatalog	DataCatalog	O catálogo para o qual este conjunto de dados pertence.

Fonte: Google (2017)

As propriedades para obtenção do conjunto de dados remetem para a localização de onde os *datasets* podem ser obtidos, seu formato e o *link* para o *download*, apresentado no quadro 4.

Quadro 4 - Propriedades para obtenção do conjunto de dados

Propriedade	Tipo	Descrição
distribution	DataDownload	Descrição do local para download do conjunto de dados e do formato de arquivo para download.
Distribution.fileFormat	Texto, recomendado	O formato de arquivo desta distribuição.
Distribution.contentUrl	URL, necessário	O link para o download.

Fonte: Google (2017)

Um conjunto de dados por ser citado em uma publicação, de modo que a adição de uma referência é particularmente útil para a descoberta e recuperação do conjunto de dados, quadro 5.

Quadro 5 - Citações e publicações

Propriedade	Tipo	Descrição
citation	Texto	Uma citação para uma publicação que descreve o conjunto de dados (por exemplo, "J.Smith", como criei um conjunto de dados incrível ", Journal of Data Science, 1966")

Fonte: Google (2017)

O conjunto de dados pelo *Schema.org*, como apresentado no quadro 5, destaca-se a descrição de uma citação, uma das formas de identificação. Destaca-se que desafio relacionado ao tratamento de dados não estruturados para extração da semântica é de grande relevância para o entendimento da recuperação da informação, bem como, o tema pode ter um viés de outras pesquisas, quanto a questão terminológica e ontológica de seus metadados.

A definição do conceito de dados não é muito clara chegando ao ponto de se definir dado como “Qualquer outra coisa que se pareça com dados” (NOY, 2017). Materializar o conceito de dados não é uma tarefa corriqueira, no entanto, percebe-se a intenção dos mantenedores da iniciativa de não restringir seu escopo.

Este também é um problema para a própria ciência. Sayão e Sales (2015, p. 18) apresentam em seu Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores uma lista não exaustiva do que pode ser considerado dados que podem ser produzidos como produto de uma pesquisa, quais sejam, dados observacionais, experimentais, brutos ou derivados, simulações, coleções físicas, modelos, software, imagens, vídeos e muito mais.

Existem também dúvidas quanto a identificação do conjunto de dados, as formas de relacionar e propagar conjuntos de dados entre si. Entretanto, o desafio técnico mais relevante para a área de Ciência da Informação está relacionada à descrição do conteúdo dos conjuntos de dados que, para os mantenedores do projeto, é crucial para a descoberta e reutilização desses

datasets, suscitando, inclusive o seguinte questionamento: "Como podemos usar eficientemente descrições de conteúdo que os provedores já descrevem de forma declarativa usando os padrões do W3C para descrever a semântica de recursos da *Web* e dados vinculados? " (NOY, 2017).

3 CONSIDERAÇÕES FINAIS

Tratando-se de uma iniciativa recém lançada e em fase de desenvolvimento, mas com grande potencial de aplicação com resultados práticos, o padrão de metadados *Schema.org* enfrenta inúmeros desafios técnicos que foram elencados pelos próprios desenvolvedores.

Apresenta-se como uma proposta inovadora e potencialmente funcional, podendo, quando implementado pelos buscadores, tornar acessíveis uma infinidade de tipos de conjuntos de dados. Seu desafio técnico mais relevante para a área de Ciência da Informação está relacionado à descrição do conteúdo dos conjuntos de dados, ou seja, seus metadados.

Description of Web datasets with Schema.org

ABSTRACT
The reuse of data contributes to minimizing the duplication of collection work, optimizing costs and resources; enabling long-term preservation while maintaining data integrity; provides safeguards against scientific misconduct, including fraud and training tools for new generations of researchers. In this way, this work is contextualized by the questioning of how to describe the datasets produced in scientific research to allow the discovery and retrieval of these sets of research data on the Web? Thus, it aims to analyze the initiative of representation of digital resources to enrich information Schema.org as an alternative to represent the data. Of qualitative approach and from the point of view of its objectives being classified as exploratory research, involving bibliographical survey and analysis of examples. Because it is a newly launched and development-based initiative with a great potential for application with practical results, the Schema.org standard for research data faces several challenges, presenting itself as an innovative and potentially functional proposal. fully implemented, make Web-based datasets accessible.
Keywords: *Enrichment of data. Schema.org. Management of scientific data*

REFERÊNCIAS

CURTY, R. G.; CERVANTES, B. M. N. Data Science: Ciência orientada a dados. **Informação & Informação**, v. 21, n. 2, p. 1–4, 20 dez. 2016.

DATA CITATION SYNTHESIS GROUP. **Joint Declaration of Data Citation Principles - FINAL**. San Diego, 30 out. 2013. Disponível em: <https://www.force11.org/group/joint-declaration-data-citation-principles-final>. Acesso em: 3 abr. 2017.

FONS, T.; PENKA, J.; WALLIS, R. OCLC´s Linked Data Initiative: Using *Schema.org* to Make Library Data Relevant on the *Web*. **Information Standards Quarterly**, Linked Data for Libraries, Archives, and Museums. v. 24, n. 2/3, Spring/Summer 2012.

GOOGLE. **Datasets | Search**. Disponível em: <https://developers.google.com/search/docs/data-types/datasets>. Acesso em: 4 nov. 2017.

HARVEY, D. R. **Digital Curation: A How-To-Do-It Manual**. New York: Neal-Schuman Publishers, 2010.

MOLLOY, J. C. The Open Knowledge Foundation: Open Data Means Better Science. **PLOS Biology**, v. 9, n. 12, p. e1001195, 6 dez. 2011.

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION, DEPARTMENT OF COMMERCE. **NCDC Storm Events Database - Data.gov**. Disponível em: <https://catalog.data.gov/dataset/ncdc-storm-events-database>. Acesso em: 26 nov. 2017.

NOY, N. **Facilitating the discovery of public datasetsGoogle Research Blog**, 24 jan. 2017. Disponível em: <https://research.googleblog.com/2017/01/facilitating-discovery-of-public.html>. Acesso em: 4 nov. 2017

SALES, L. F.; CAVALCANTI, M. T. SELEÇÃO E AVALIAÇÃO DE COLEÇÕES DE DADOS DIGITAIS DE PESQUISA: uma possível abordagem metodológica. **Informação & Tecnologia**, v. 2, n. 2, p. 88–105, 2015.

SANTOS, P. L. V. A. DA C.; SIMIONATO, A. C.; ARAKAKI, F. A. Definição de metadados para recursos informacionais: apresentação da metodologia BEAM. **Informação & Informação**, v. 19, n. 1, p. 146, 25 fev. 2014.

SAYÃO, L. F.; SALES, L. F. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 8, n. 2, 2013.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores.** [s.l.] Instituto de Engenharia Nuclear, 2015.

SAYÃO, L. F.; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ: novas práticas em informação e conhecimento**, v. 5, n. 2, p. 67–71, 9 jan. 2017.

SCHEMA.ORG. about page - schema.org. Disponível em: <https://schema.org/docs/about.html>.

STATISTA. **Brazil search engine market share 2017 | Statistic**. Disponível em: <https://www.statista.com/statistics/309652/brazil-market-share-search-engine/>. Acesso em: 26 nov. 2017.