

Boas Práticas para Publicação de Dados na Web

Bernadette Lóscio, Caroline Burle and Newton Calegari



ceweb.br **nic.br** **cgi.br**

Quem somos



Newton Bernadette Caroline

Tópicos a serem discutidos

- Contexto de Dados na Web
 - Dados na Web X Dados Abertos x Dados Conectados
- Casos de uso de Dados na Web
 - Desafios e Requerimentos de Dados na Web
- Boas Práticas para Publicação de Dados na Web
 - Benefícios das Boas Práticas para Publicação de Dados na Web

Como facilitar o reuso dos dados?

É fundamental que haja uma compreensão mútua entre publicadores e consumidores dos dados.

Sem esse entendimento, o esforço dos publicadores pode ser incompatível com o desejo dos consumidores.



Consome dados

Boas
Práticas



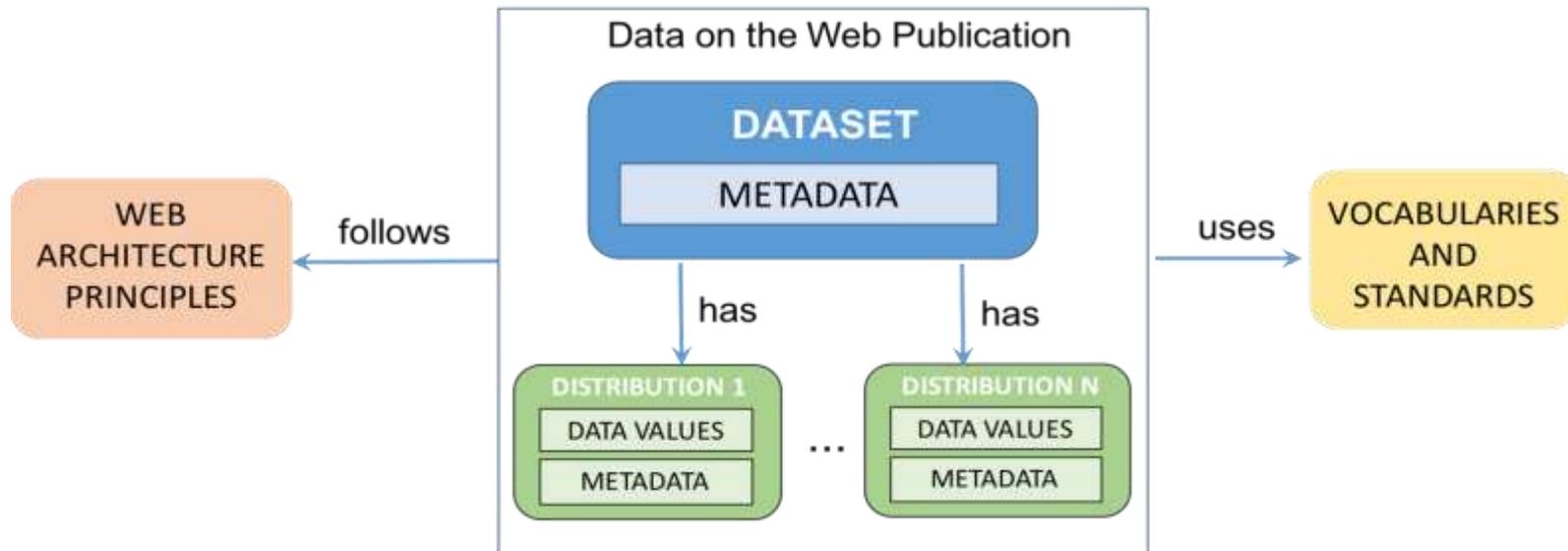
Publica dados

Grupo de Trabalho de Boas Práticas para Publicação de Dados na Web

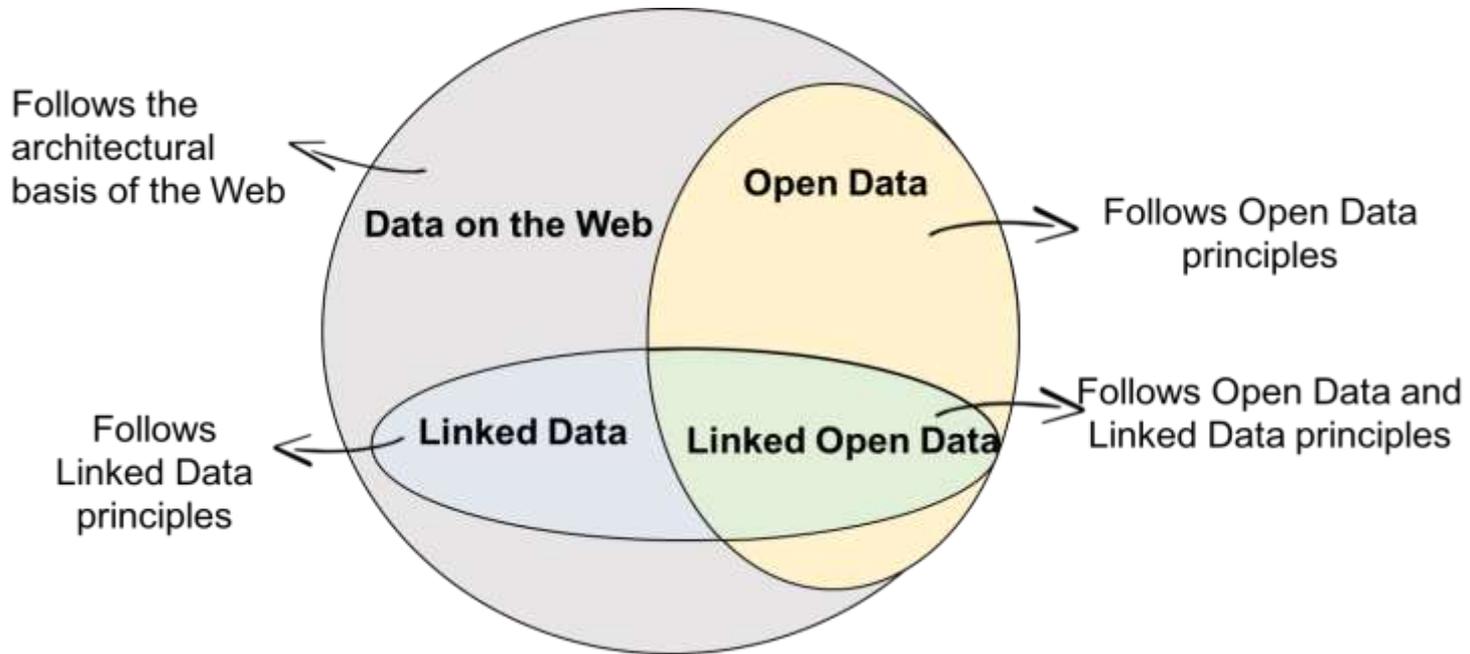
A **Missão** do Grupo de Trabalho de Boas Práticas para Publicação de Dados na Web, parte do [Data Activity](#), do W3C, é:

1. Desenvolver o **ecossistema de dados abertos**, facilitando a comunicação entre desenvolvedores e publicadores de dados;
2. Providenciar **diretrizes para os publicadores de dados**, que promoverá consistência na forma como os dados são gerenciados, provendo assim o reuso de dados;
3. **Fomentar a confiança dos desenvolvedores nos dados**, independente da tecnologia utilizada, aumentando o potencial de inovação.

Contexto dos Dados na Web



Dados na Web x Dados Abertos x Dados Conectados



Casos de Uso de Dados na Web



Data on the Web Best Practices Use Cases & Requirements

W3C Working Group Note 24 February 2015

This version:

<http://www.w3.org/TR/2015/NOTE-dwbp-ucr-20150224/>

Latest published version:

<http://www.w3.org/TR/dwbp-ucr/>

Latest editor's draft:

<http://w3c.github.io/dwbp/usecasesv1.html>

Previous version:

<http://www.w3.org/TR/2014/WD-dwbp-ucr-20141014/>

Editors:

[Deirdre Lee](#), [Derilinx](#) (formerly at Insight@NUIG, Ireland)

[Bernadette Farias Lôscio](#), [Centro de Informática - Universidade Federal de Pernambuco, Brazil](#)

[Phil Archer](#), [W3C/ERCIM](#)

<https://www.w3.org/TR/dwbp-ucr/>

Casos de Uso de Dados na Web

1. Introduction
2. Use Cases
 - 2.1 ASO: Airborne Snow Observatory
 - 2.2 BBC
 - 2.3 Bio2RDF
 - 2.4 BuildingEye: SME use of public data
 - 2.5 Dados.gov.br
 - 2.6 Digital archiving of Linked Data
 - 2.7 Dutch Base Registers
 - 2.8 GS1 Digital
 - 2.9 ISO GEO Story
 - 2.10 The Land Portal
 - 2.11 LA Times' Reporting of Ron Galperin's Infographic
 - 2.12 LusTRE: Linked Thesaurus fRamework for Environment
 - 2.13 Machine-readability and Interoperability of Licenses
 - 2.14 Mass Spectrometry Imaging (MSI)
 - 2.15 OKFN Transport WG
 - 2.16 Open City Data Pipeline
 - 2.17 Open Experimental Field Studies
 - 2.18 Resource Discovery for Extreme Scale Collaboration (RDESC)
 - 2.19 Recife Open Data Portal
 - 2.20 Retrato da Violência (Violence Map)
 - 2.21 Share-PSI 2.0: Uses of Open Data Within Government for Innovation and Efficiency
 - 2.22 Tabulae - how to get value out of data
 - 2.23 UK Open Research Data Forum
 - 2.24 Uruguay Open Data Catalog
 - 2.25 Web Observatory
 - 2.26 Wind Characterization Scientific Study
3. General Challenges
 - 3.1 A Word on Open and Closed Data
 - 3.2 Requirements by Challenge
4. Requirements
 - 4.1 Requirements for Data on the Web Best Practices
 - 4.2 Requirements for Quality and Granularity Description Vocabulary
 - 4.3 Requirements for Data Usage Description Vocabulary

*12 desafios e
42 requerimentos*

Publicando Dados na Web

*Como
disponibilizar
os dados?*

*Quais dados
publicar?*

*Como deixar os
dados
interoperáveis?*

*Quais são as
fontes dos
dados?*

*Como identificar
os recursos dos
dados?*

*Quais
formatos
utilizar?*

*Como
receber
feedback?*

*Publicar dados na Web é mais do que
apenas publicar dados!*

Desafios para Publicar Dados na Web

- Metadados (*para humanos e máquinas*)
- Licença de Dados (*como permitir & restringir o acesso?*)
- Proveniência & Qualidade dos Dados (*como adicionar confiança?*)
- Versionamento dos Dados (*rastrear as versões dos conjuntos de dados*)
- Identificação dos Dados (*identificação dos conjuntos de dados e distribuições*)
- Formatos dos Dados (*quais formatos usar?*)

Desafios para Publicar Dados na Web

- Vocabulários dos Dados (*como promover a interoperabilidade?*)
- Acesso aos Dados (*opções de acesso*)
- Preservação dos Dados
- Feedback (*como engajar usuários?*)
- Enriquecimento dos Dados (*adicionar valor aos dados*)
- Republicação dos Dados (*responsabilidade de reusar os dados*)

Data on the Web Best Practices

W3C Candidate Recommendation 30 August 2016



This version:

<https://www.w3.org/TR/2016/CR-dwbp-20160830/>

Latest published version:

<https://www.w3.org/TR/dwbp/>

Latest editor's draft:

<http://w3c.github.io/dwbp/bp.html>

Implementation report:

https://www.w3.org/2013/dwbp/wiki/BP_Implementation_Report

Previous version:

<http://www.w3.org/TR/2016/WD-dwbp-20160519/>

Editors:

Bernadette Farias Lóscio, [CIn - UFPE, Brazil](#)

Caroline Burle, [NIC.br, Brazil](#)

Newton Calegari, [NIC.br, Brazil](#)

Contributors:

Annette Greiner

Antoine Isaac

Carlos Iglesias

Carlos Laufer

Christophe Guéret

Deirdre Lee

Eric G. Stephan

Eric Kauz

Ghislain A. Atemezing

Hadley Beeman

<https://www.w3.org/TR/dwbp/>

[Best Practice 1](#): Provide metadata

[Best Practice 2](#): Provide descriptive metadata

[Best Practice 3](#): Provide structural metadata

[Best Practice 4](#): Provide data license information

[Best Practice 5](#): Provide data provenance information

[Best Practice 6](#): Provide data quality information

[Best Practice 19](#): Use content negotiation for serving data available in multiple formats

Evidence

Relevant requirements: [R-ProvAvailable](#), [R-MetadataAvailable](#)

[Best Practice 23](#): Make data available through an API

Intended Outcome

Humans will know the origin or history of the dataset and software agents will be able to automatically process provenance information.

[Best Practice 10](#): Use persistent UHIs as identifiers within datasets

[Best Practice 11](#): Assign URIs to dataset versions and series

[Best Practice 12](#): Use machine-readable standardized data formats

[Best Practice 13](#): Use locale-neutral data representations

[Best Practice 14](#): Provide data in multiple formats

[Best Practice 15](#): Reuse vocabularies, preferably standardized ones

[Best Practice 16](#): Choose the right formalization level

[Best Practice 17](#): Provide bulk download

[Best Practice 18](#): Provide Subsets for Large Datasets

[Best Practice 26](#): Avoid Breaking Changes to Your API

[Best Practice 27](#): Preserve identifiers

[Best Practice 28](#): Assess dataset coverage

[Best Practice 29](#): Gather feedback from data consumers

[Best Practice 30](#): Make feedback available

[Best Practice 31](#): Enrich data by generating new data

[Best Practice 32](#): Provide Complementary Presentations

[Best Practice 33](#): Provide Feedback to the Original Publisher

[Best Practice 34](#): Follow Licensing Terms

[Best Practice 35](#): Cite the Original Publication

Provide metadata that describes the overall features of datasets and distributions.

Why

Explicitly providing dataset descriptive information allows user agents to automatically discover datasets available on the Web and it allows humans to understand the nature of the dataset and its distributions.

Intended Outcome

Humans will be able to interpret the nature of the dataset and its distributions, and software agents will be able to automatically discover datasets and distributions.

Possible Approach to Implementation

Descriptive metadata can include the following overall features of a dataset:

- The **title** and a **description** of the dataset.
- The **keywords** describing the dataset.
- The **date of publication** of the dataset.
- The **entity responsible (publisher)** for making the dataset available.
- The **contact point** for the dataset.
- The **spatial coverage** of the dataset.
- The **temporal period** that the dataset covers.
- The **date of last modification** of the dataset.
- The **themes/categories** covered by a dataset.

Descriptive metadata can include the following overall features of a distribution:

- The **title** and a **description** of the distribution.
- The **date of publication** of the distribution.
- The **media type** of the distribution.

The machine-readable version of the descriptive metadata can be provided using the vocabulary recommended by W3C to describe datasets, i.e. the Data Catalog Vocabulary [[VOCAB-DCAT](#)]. This provides a framework in which datasets can be described as abstract entities.

Machine-readable metadata

EXAMPLE 2

Machine-readable

The example below shows how to use [\[VOCAB-DCAT\]](#) to provide the machine-readable **discovery** metadata for the bus stops dataset ([stops-2015-05-05](#)). The dataset has one CSV distribution ([stops-2015-05-05.csv](#)) that is also described using the [\[VOCAB-DCAT\]](#). The dataset is classified under the domain represented by the relative URI [mobility](#). This domain may be defined as part of a set of domains identified by the URI [themes](#). To describe both concepts and schema concepts, John used [SKOS](#). To express frequency of update an instance from the [Content-Oriented Guidelines](#) developed as part of the W3C Data Cube Vocabulary efforts was used. John chose to describe the spatial and temporal coverage of the example dataset using URIs from [Geonames](#) and the [Interval dataset](#) from data.gov.uk, respectively.

```
:stops-2015-05-05
  a dcat:Dataset ;
  dct:title "Bus stops of MyCity" ;
  dcat:keyword "transport","mobility","bus" ;
  dct:issued "2015-05-05"^^xsd:date ;
  dcat:contactPoint <http://data.mycity.example.com/transport/contact> ;
  dct:temporal <http://reference.data.gov.uk/id/year/2015> ;
  dct:spatial <http://www.geonames.org/3399415> ;
  dct:publisher :transport-agency-mycity ;
  dct:accrualPeriodicity <http://purl.org/linked-data/sdmx/2009/code#freq-A>
  dcat:theme :mobility ;
  dcat:distribution :stops-2015-05-05.csv ;
```

Human-readable metadata

Dataset description

Title	Bus timetable of MyCity
URI	http://data.mycity.example.com/transport/dataset/bus/stops-2015-05-05
Keywords	transport, mobility, bus
Publication date	2015-05-05
Publisher	Transport Agency MyCity
Creator	John < john@mycitytransport.org >
Contact point	http://data.mycity.example.com/transport/contact
Period that the dataset covers	The British calendar year of 2014
Spatial coverage	Fortaleza, Brazil
Update frequency	Annual
Theme	Mobility
Language	English, Portuguese
Date and time formats	ISO 8601
Current version	1.2

Benefícios das Boas Práticas para Publicação de Dados na Web

Cada benefício representa uma melhoria na maneira como os conjuntos de dados são disponibilizados na Web



Reuse



Comprehension



Linkability



Discoverability



Trust



Access



Interoperability



Processability

Reuse

- BP: Provide data license information
- BP: Provide versioning information
- BP: Provide version history
- BP: Use non-proprietary data formats
- BP: Provide data in multiple formats
- BP: Use a trusted serialization format for preserved data dumps
- BP: Enrich data by generating new metadata
- BP: Provide data provenance information
- BP: Provide data quality information
- BP: Use persistent URIs as identifiers

Trustworthy

- BP: Assess dataset coverage
- BP: Assign URIs to dataset versions and series
- BP: Provide data up to date
- BP: Update the status of identifiers
- BP: Gather feedback from data consumers
- BP: Provide information about feedback
- BP: Provide data provenance information
- BP: Provide data quality information

Comprehension

- BP: Provide metadata
- BP: Provide locale parameters metadata
- BP: Provide structural metadata
- BP: Provide descriptive metadata

Accessibility

- BP: Provide bulk download
- BP: Follow REST principles when designing APIs
- BP: Provide real-time access
- BP: Maintain separate versions for a data API
- BP: Assess dataset coverage

Discoverability

- BP: Provide descriptive metadata
- BP: Use persistent URIs as identifiers
- BP: Assign URIs to dataset versions and series

Linkability

- BP: Use persistent URIs as identifiers
- BP: Assign URIs to dataset versions and series

Processability

- BP: Use machine-readable standardized data formats
- BP: Enrich data by generating new metadata

Interoperability

- BP: Use standardized terms
- BP: Re-use vocabularies

Best Practice 1: Provide metadata

Metadata must be provided for both human users and computer applications

Why

Providing metadata is a fundamental requirement which publishers and data consumers may be unknown to each other. Metadata that helps human users and computer applications to understand aspects that describes a dataset or a distribution.

Intended Outcome

Human-readable metadata will enable humans to understand the data. Machine-readable metadata will enable computer applications, notably

Possible Approach to Implementation

Possible approaches to provide *human readable metadata*

- to provide metadata as part of an HTML Web page
- to provide metadata as a separate text file

Possible approaches to provide *machine readable metadata*

- machine readable metadata may be provided in HTML. If it can be embedded in the HTML page using [HTML metadata](#), it can be published separately, they should be served from a single source. The coexistence of multiple formats is best achieved by generating a single source of the metadata.
- when defining machine readable metadata, reuse existing vocabularies are strongly recommended. For example, Dublin Core and Data Catalog Vocabulary [\[VOCAB-DCAT\]](#) should be used.

Benefícios das Boas Práticas

- **Compreensão:** humanos terão um melhor entendimento sobre a estrutura dos dados, seu significado, os metadados e a natureza do conjunto de dados.
- **Processamento:** máquinas poderão automatizar o processo e manipular os dados do conjunto de dados.
- **Descobrimento:** máquinas poderão descobrir automaticamente os dados ou um conjunto de dados.
- **Reuso:** possibilidade de aumentar o reuso de conjuntos de dados por distintos grupos de consumidores.

Chamada para **Implementação** das Boas Práticas para Publicação de Dados na Web

DWBP Evidences Form

Thank you for your help to collect implementation evidence of the Data on the Web Best Practices (DWBP), a document to be released as a W3C Recommendation in 2016.

There are 35 forms, each one of them is referring to one Best Practice and there is the possibility to create more evidences for each BP.

The Data on the Web Best Practices document is available at <https://www.w3.org/TR/dwbp>.

The Organisation name will be mentioned as one of the organisations that tested the **Data on the Web Best Practices** in order to become a W3C Recommendation.

You may want to start filling your name and email and then the name of the organisation that published the resource or dataset. After that please start filling the respective forms for the best practices that were implemented.

If you have some questions, feel free to send us a message (public-dwbp-comments@w3.org).

Contact information

Name

Email

Publisher's information (Organisation that published the resource or dataset)

Publisher

Site

Save info

<http://w3c.br/form-dwbp/>

Obrigada!

www.ceweb.br - www.cin.ufpe.br



cburle@nic.br



bfl@cin.ufpe.br



newton@nic.br



@carolburle



@bernafarias



@newtoncalegari

São Paulo, Brasil

14/10/2016