Data on the Web Best Practices: Challenges and Benefits

Bernadette Lóscio, Caroline Burle and Newton Calegari
Topics to be discussed

• Data on the Web Context
• Data on the Web use cases
• Data on the Web Challenges and Requirements
• Data on the Web Best Practices
• Data on the Web Best Practices Benefits
Data on the Web x Open Data x Linked Data

Follows the architectural basis of the Web

Follows Open Data principles

Follows Open Data and Linked Data principles
Data on the Web Context

Data on the Web Publication

DATASET

METADATA

DISTRIBUTION 1
DATA VALUES
METADATA

DISTRIBUTION N
DATA VALUES
METADATA

WEB ARCHITECTURE PRINCIPLES
follows

VOCABULARIES AND STANDARDS
uses
Players of the data on the Web ecosystem

Data publisher: publishes and shares data

Data consumer: reuses the data and might generate new data

Several types of data sources (transactional systems, sensors, mobile devices, documents...)

How to enable the data reuse?

Source: http://ceweb.br/livros/dados-abertos-conectados/capitulo-1/
How to enable the data reuse?

A common understanding between data publishers and data consumers becomes fundamental. Without this agreement, data publishers' efforts may be incompatible with data consumers' desires.
The **Mission** of the Data on the Web Best Practices Working Group, part of the [Data Activity](https://www.w3.org/2013/dwbp/wiki/Main_Page), is:

1. to develop the **open data ecosystem**, facilitating better communication between developers and publishers;
2. to provide **guidance to publishers** that will improve consistency in the way data is managed, thus **promoting the re-use of data**;
3. to **foster trust in the data** among developers, whatever technology they choose to use, increasing the potential for **genuine innovation**.
Data on the Web use cases

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 Lóscio, Centro de Informática - Universidade Federal de Pernambuco, Brazil
Phil Archer, W3C/ERCIM

https://www.w3.org/TR/dwbp-ucr/
Publishing data on the Web is more than just publishing data!

- Which data to publish?
- Which data sources?
- Which data formats to use?
- How to make data available?
- How to make data interoperable?
- How to identify data resources?
- How to gather feedback?
Data on the Web Challenges

- Metadata *(for humans & machines)*
- Data Licenses *(how to permit & restrict access?)*
- Data Provenance & Quality *(how to add trust?)*
- Data Versioning *(tracking dataset versions)*
- Data Identification *(identifying datasets and distributions)*
- Data Formats *(which data formats to use?)*
Data on the Web Challenges

- Data Vocabularies *(how to promote interoperability?)*
- Data Access *(access options)*
- Data Preservation
- Feedback *(how to engage users?)*
- Data Enrichment *(adding value to data)*
- Data Republication *(reuse data responsibly)*

*12 challenges and 42 requirements*
Data on the Web Best Practices

W3C Recommendation 31 January 2017

This version:
https://www.w3.org/TR/2017/REC-dwbp-20170131/

Latest published version:
https://www.w3.org/TR/dwbp

Latest editor's draft:
http://w3c.github.io/dwbp/bp.html

Implementation report:
http://w3c.github.io/dwbp/dwbp-implementation-report.html

Previous version:
https://www.w3.org/TR/2016/PR-dwbp-20161215/

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 Lauffer
Christophe Guéret
Deirdre Lee
Doug Schapers
Eric G. Stephan
Eric Kauz
Ghislain A. Atemezing
Hadley Beeman
Igbert Bittencourt
Iofs Paulo Almeida

Audience:
BP are designed to meet the needs of information management staff, developers, and wider groups such as scientists interested in sharing and reusing research data on the Web

Source: http://w3c.github.io/dwbp/bp.html
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 7: Provide data in multiple formats
Best Practice 8: Reuse vocabularies, preferably standardized ones
Best Practice 9: Choose the right formalization level
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 19: Use content negotiation for serving data available in multiple formats
Best Practice 20: Provide complementary presentations
Best Practice 21: Provide Feedback to the Original Publisher
Best Practice 22: Follow Licensing Terms
Best Practice 23: Make data available through an API

Evidence

Relevant requirements: R-ProvAvailable, R-MetadataAvailable

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 24: Use persistent UHIs as identifiers within datasets
Best Practice 25: Assign URIs to dataset versions and series
Best Practice 26: Use machine-readable standardized data formats
Best Practice 27: Use locale-neutral data representations
Best Practice 28: Provide data in multiple formats
Best Practice 29: Reuse vocabularies, preferably standardized ones
Best Practice 30: Choose the right formalization level
Best Practice 31: Provide bulk download
Best Practice 32: Provide Subsets for Large Datasets
DWBP Benefits

Each benefit represents an improvement in the way how datasets are available on the Web

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

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

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

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

- **Comprehension**: humans will have a better understanding about the data structure, the data meaning, the metadata and the nature of the dataset.

- **Processability**: machines will be able to automatically process and manipulate the data within a dataset.

- **Discoverability**: machines will be able to automatically discover a dataset or data within a dataset.

- **Reuse**: the chances of dataset reuse by different groups of data consumers will increase.
BP Benefits

- **Linkability**: it will be possible to create links between data resources (datasets and data items).
- **Interoperability**: it will be easier to reach consensus among data publishers and consumers.
- **Trust**: the confidence that consumers have in the dataset will improve.
- **Access**: humans and machines will be able to access up to date data in a variety of forms.
How can you contribute now?

TABLE OF CONTENTS
1. Introduction
   1.1 Methodology
   1.2 Meeting the exit criteria
2. DWBP Evidence
   2.1 Datasets, Data portals and Vocabularies
   2.2 Documents and References
   2.3 Guidelines
3. General analysis
4. DWBP and Data Catalogs
5. Set of Best Practices
6. Acknowledgements

DWBP Implementation Report
W3C Document 29 January 2017

Editors:
Bernadette Farias Lóscio, CIin - UFPE, Brazil
Caroline Burio, NIC.br, Brazil
Newton Calegari, NIC.br, Brazil

Abstract
This document reports on evidence and implementations of the Data on the Web Best Practices Candidate Recommendation. In particular, it demonstrates that the DWBP are already in use and are also implementable.

Status of This Document
This document is merely a W3C-internal document. It has no official standing of any kind and does not represent consensus of the W3C Membership.

1. Introduction
One of the main goals of the Data on the Web Best Practices (DWBP) is to facilitate interaction between publishers and consumers of data on the Web. A set of 35 Best Practices were created to cover different challenges related to data publishing and consumption, such as Metadata, Data licenses, Data provenance, Data quality, Data versioning, Data identification, Data formats, Data vocabularies, Data access and APIs, Data preservation, Feedback, Data enrichment and Data republication.

Fonte: http://w3c.github.io/dwbp/dwbp-implementation-report.html

Data on the Web Best Practices: Challenges and Benefits
Obrigada(o)!

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

@cburle@nic.br    @carolburle
@bfl@cin.ufpe.br  @bernafarias
@newton@nic.br    @newtoncalegari