Data on the Web Best Practices: Challenges and Benefits

Bernadette Lóscio, Caroline Burle and Newton Calegari

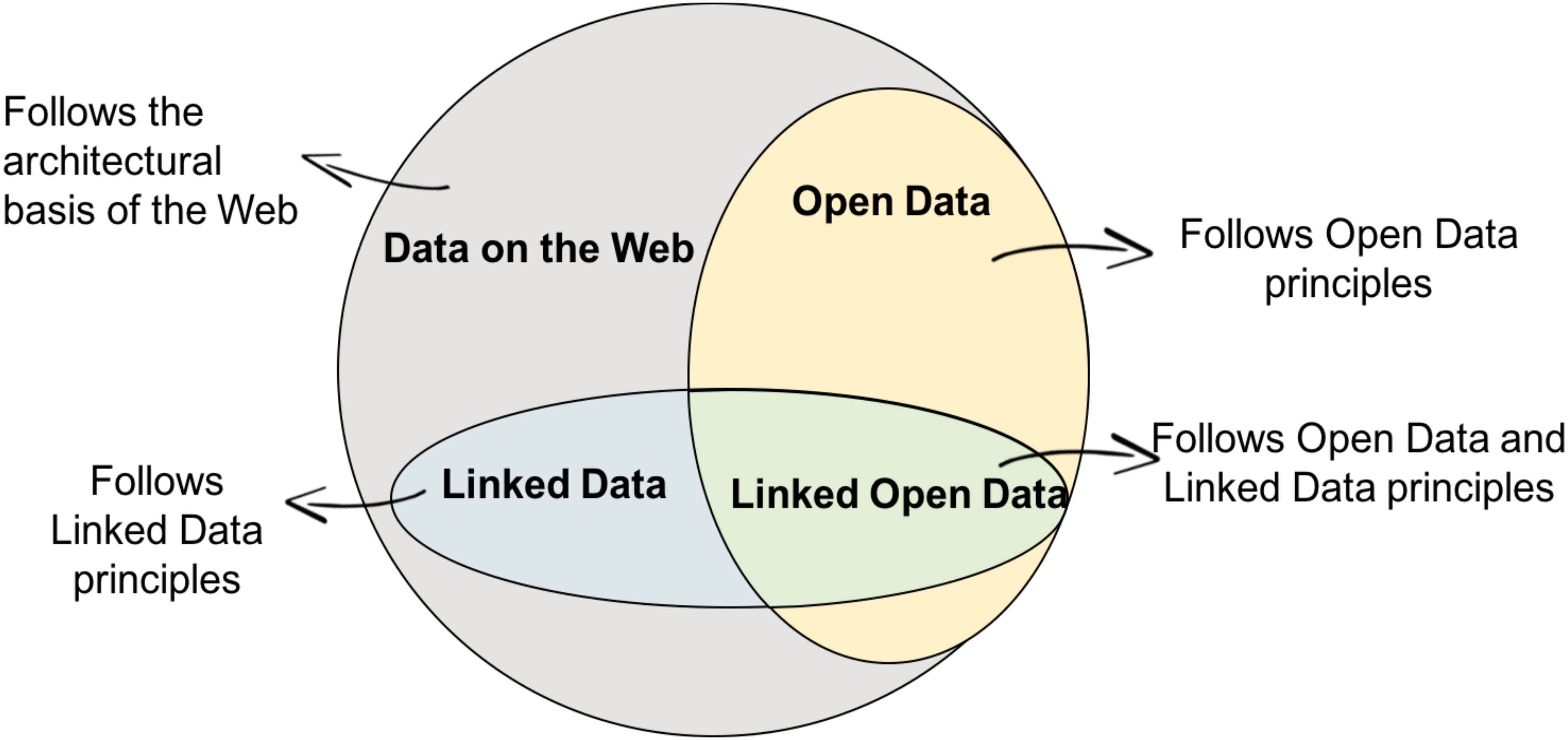
Centro de Informática U·F·P·E

ceweb.br
Centro de Estudos sobre Tecnologias Web

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

Data on the Web

Open Data

Follows Open Data principles

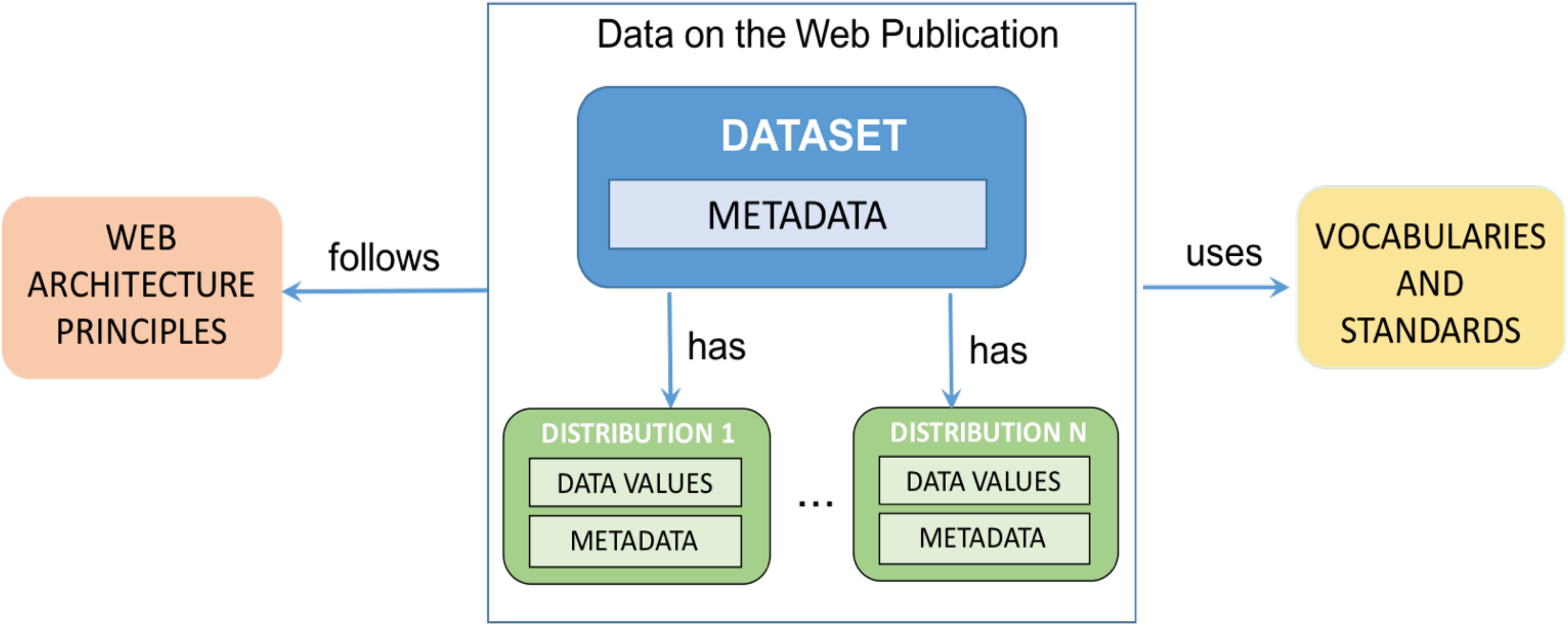
Linked Data

Linked Open Data

Follows Linked Data principles

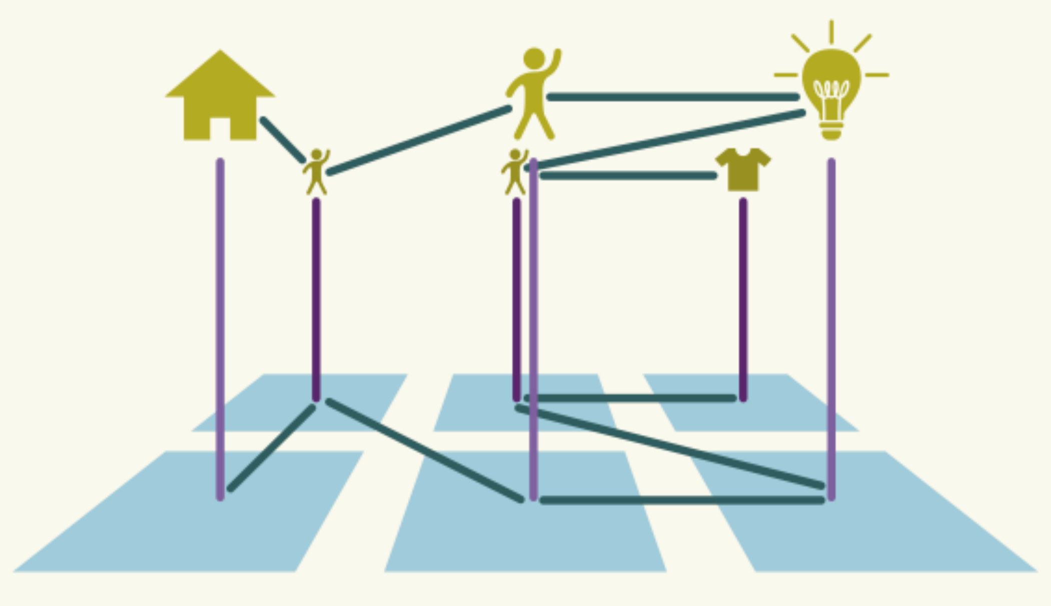
Follows Open Data and Linked Data principles

Data on the Web Context



Players of the data on the Web ecosystem

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



Source: http://ceweb.br/livros/dados-abertos-conectados/capitulo-1/

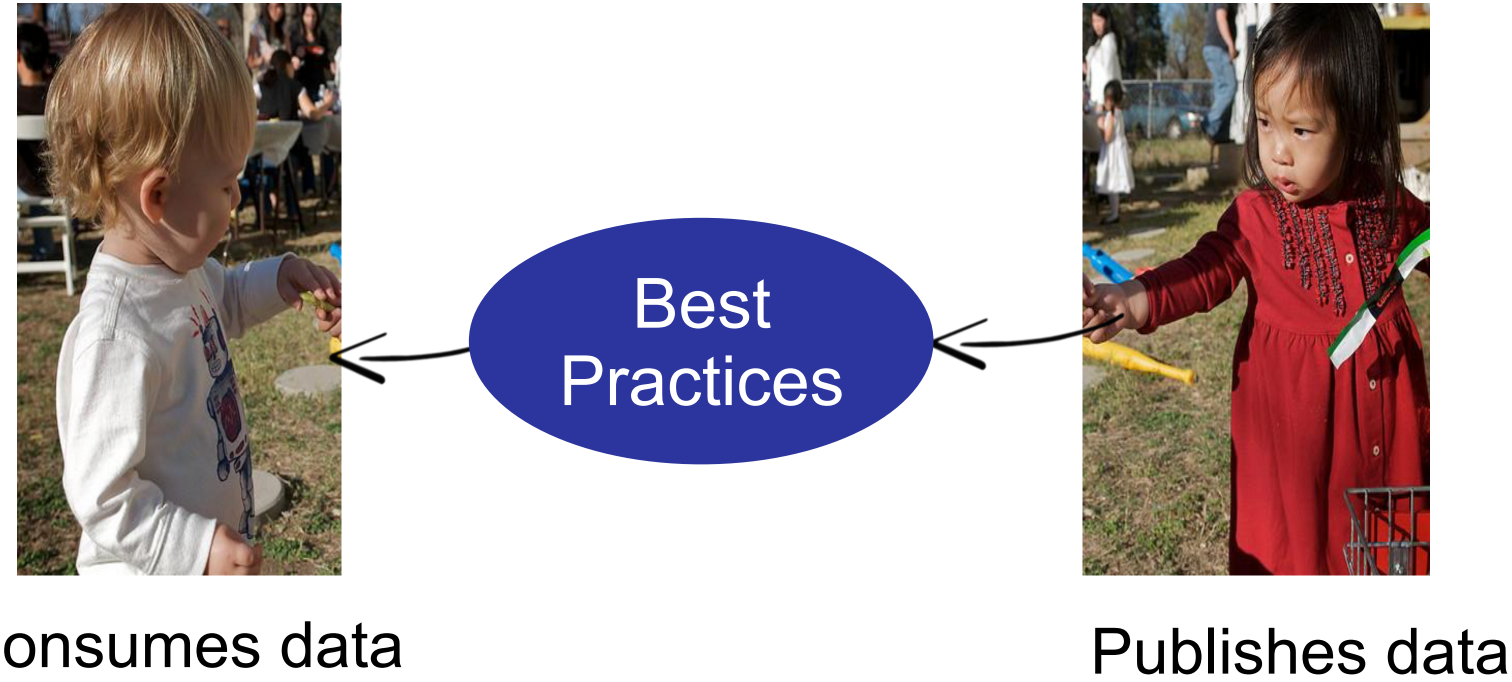
Data publisher: publishes and shares data

Data consumer: reuses the data and might generate new data

How to enable the data reuse?

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.



Best Practices

Consumes data

Publishes data

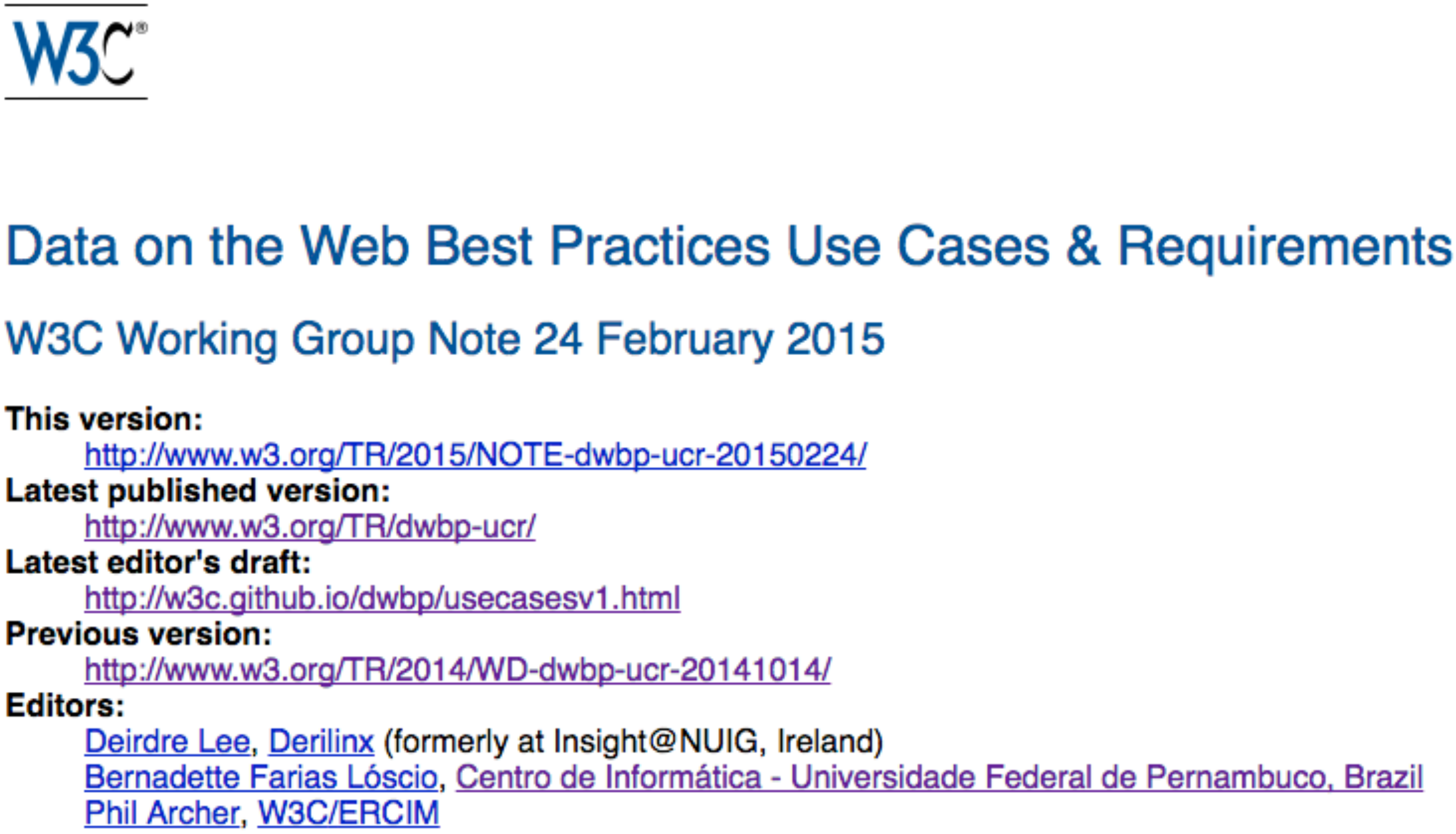
Data on the Web Best Practices Working Group

The **Mission** of the Data on the Web Best Practices Working Group, part of the Data Activity, 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.

Source: https://www.w3.org/2013/dwbp/wiki/Main_Page:

Data on the Web use cases



W3C®

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/

Data on the Web use cases

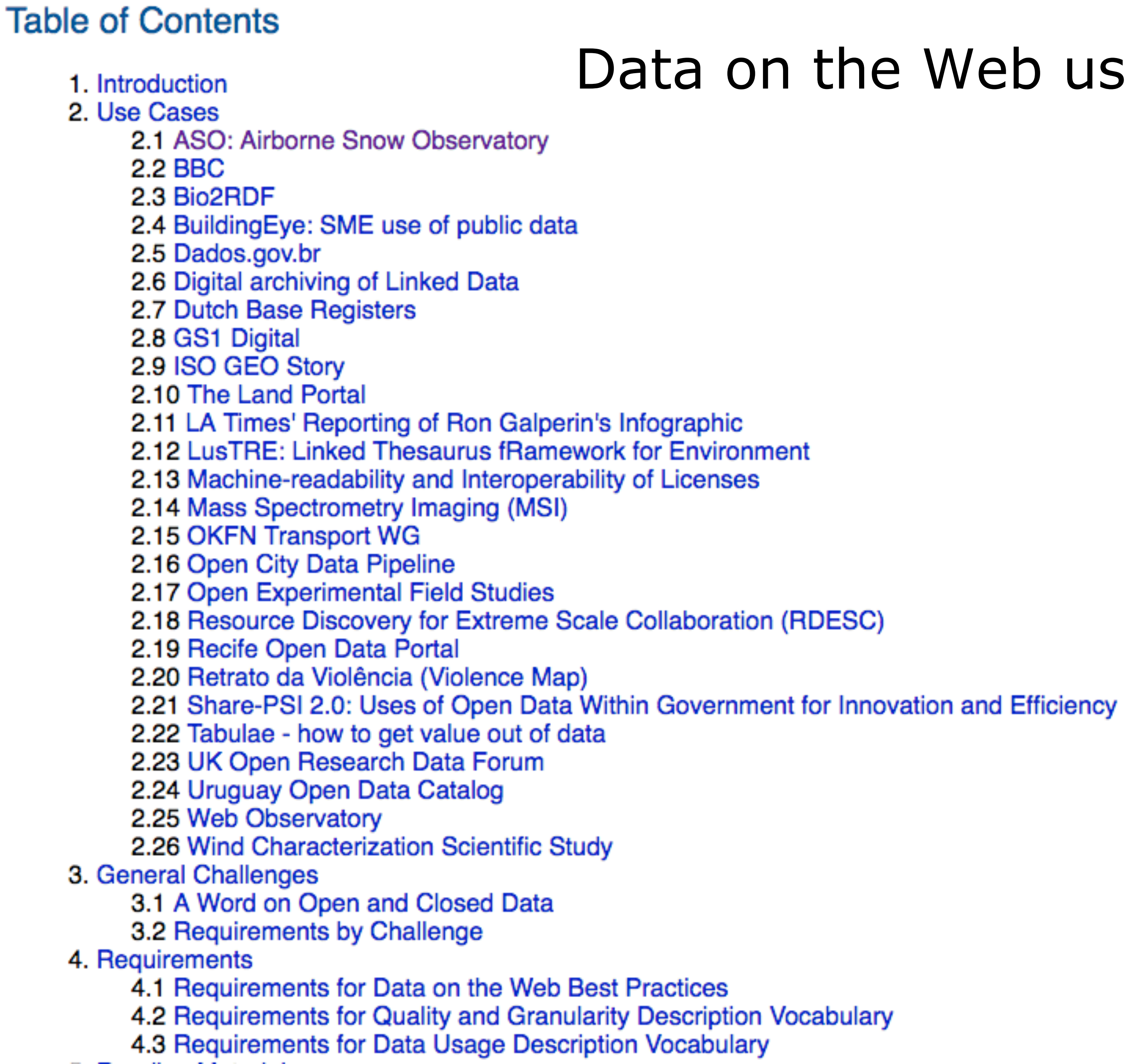
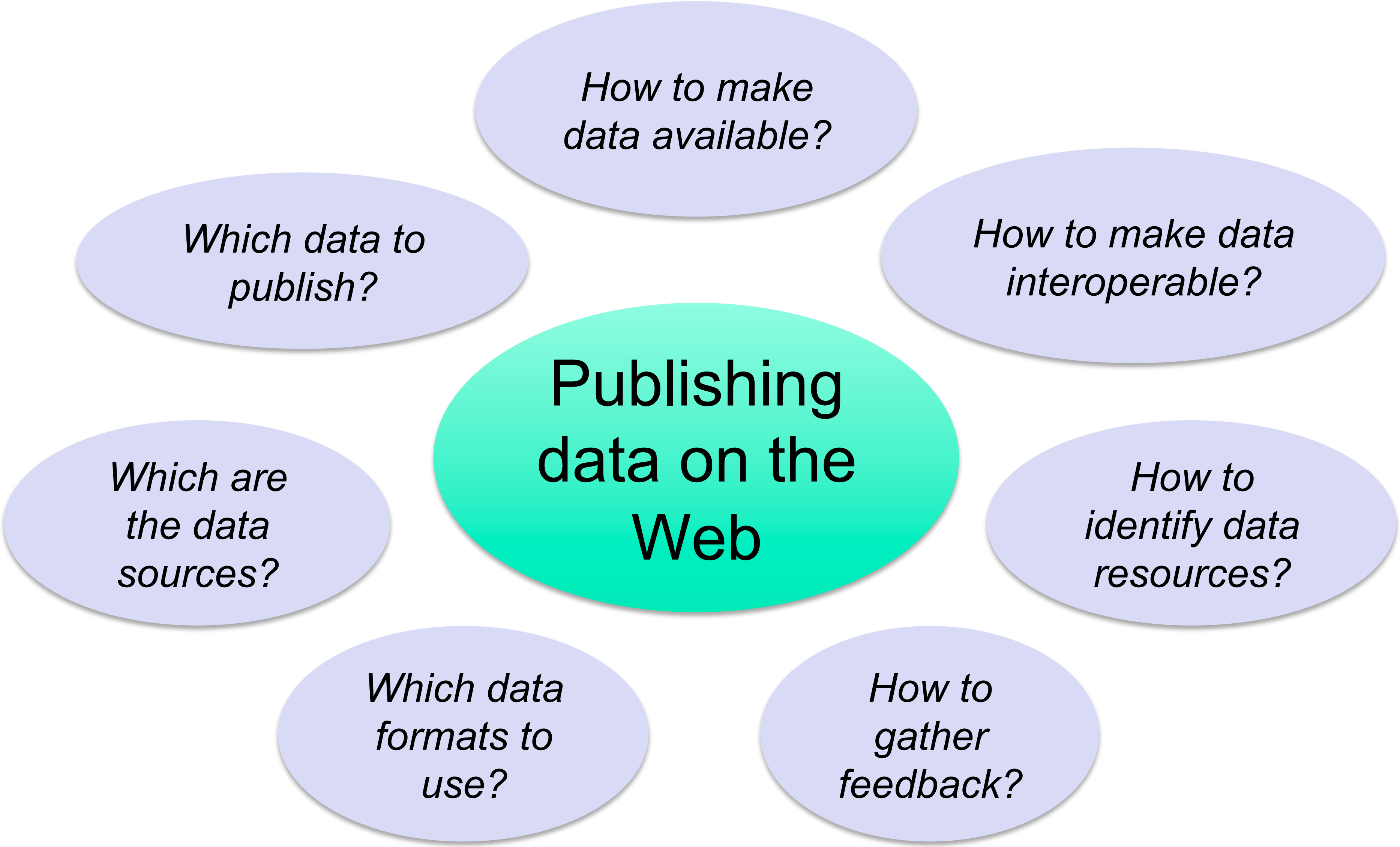


Table of Contents



Publishing data on the Web

How to make data available?

How to make data interoperable?

How to identify data resources?

How to gather feedback?

Which data formats to use?

Which are the data sources?

Which data to publish?

Publishing data on the Web is more than just publishing data!

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

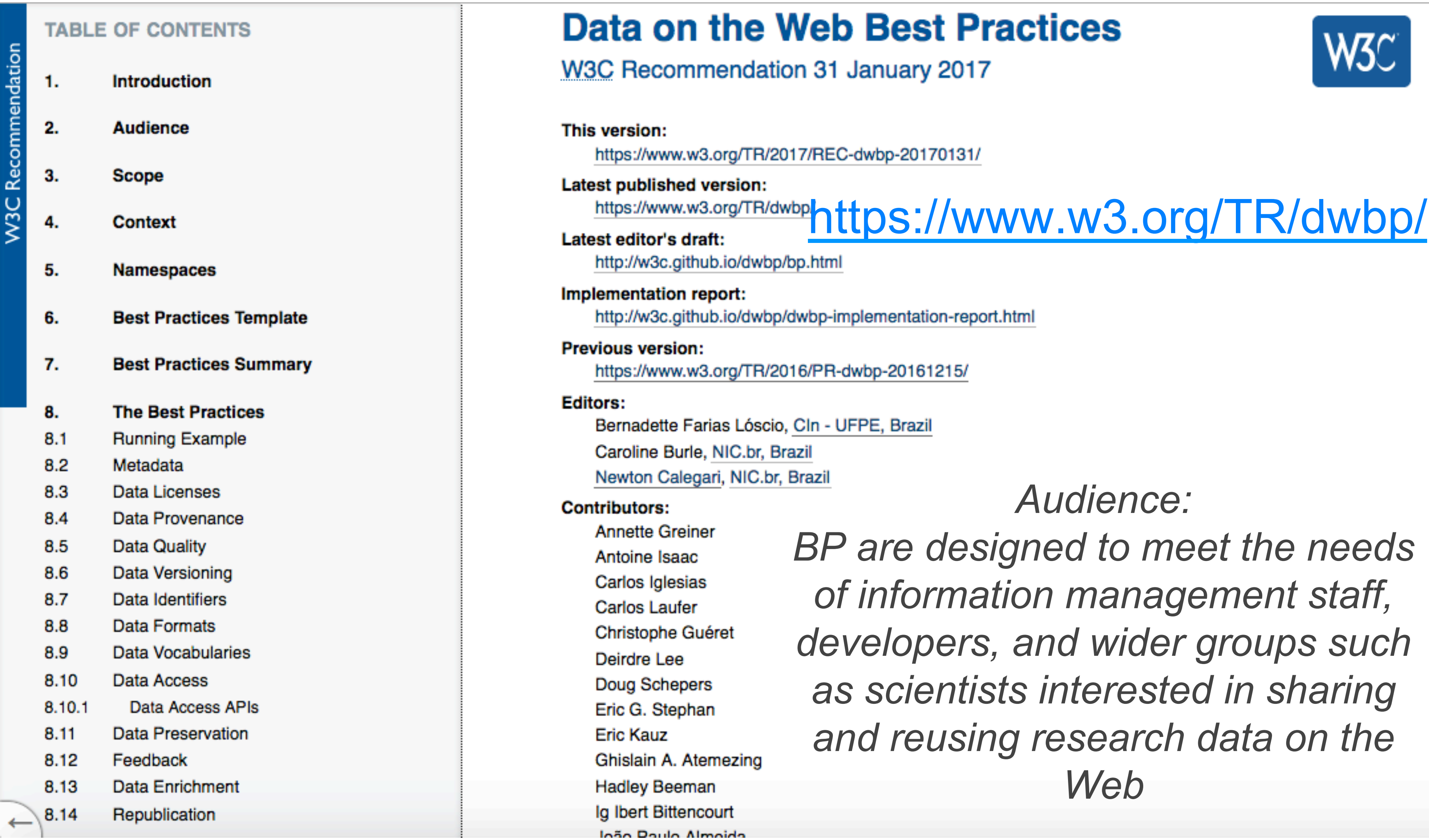


TABLE OF CONTENTS

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/

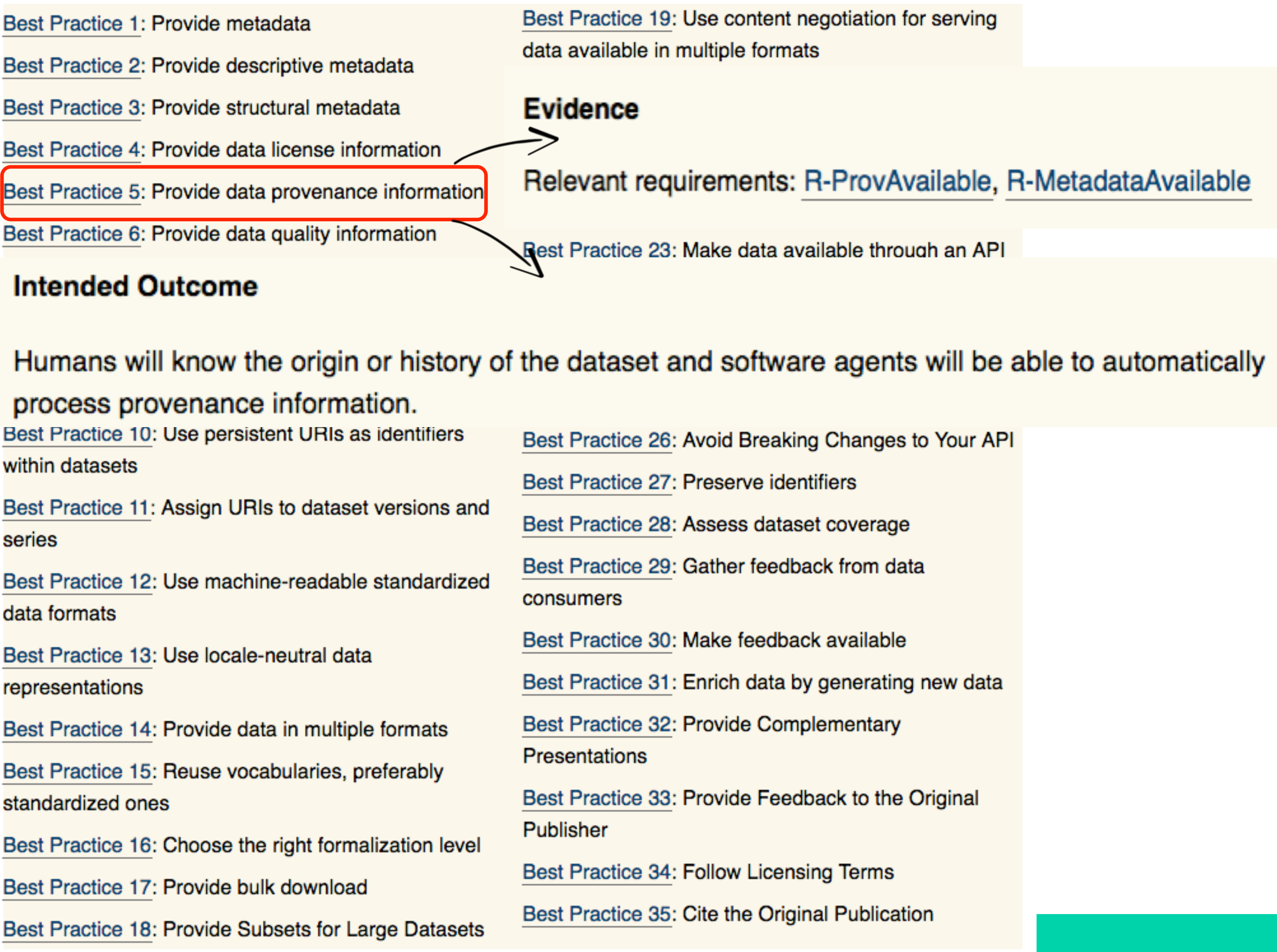
Editors:
Bernadette Farias Lóscio, CIn - UFPE, Brazil
Caroline Burle, NIC.br, Brazil
Newton Calegari, NIC.br, Brazil

Contributors:
Annette Greiner
Antoine Isaac
Carlos Iglesias
Carlos Laufer
Christophe Guéret
Deirdre Lee
Doug Schepers
Eric G. Stephan
Eric Kauz
Ghislain A. Atemezing
Hadley Beeman
Ig Ibert Bittencourt
João Paulo Almeida

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

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

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 URIs 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

Evidence

Relevant requirements: R-ProvAvailable, R-MetadataAvailable

Best Practice 23: Make data available through an API

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

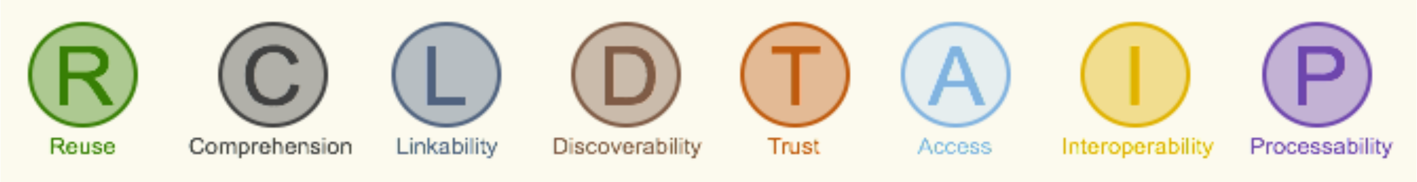
Best Practice 33: Provide Feedback to the Original Publisher

Best Practice 34: Follow Licensing Terms

Best Practice 35: Cite the Original Publication

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

Discoverability

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

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

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

Comprehension

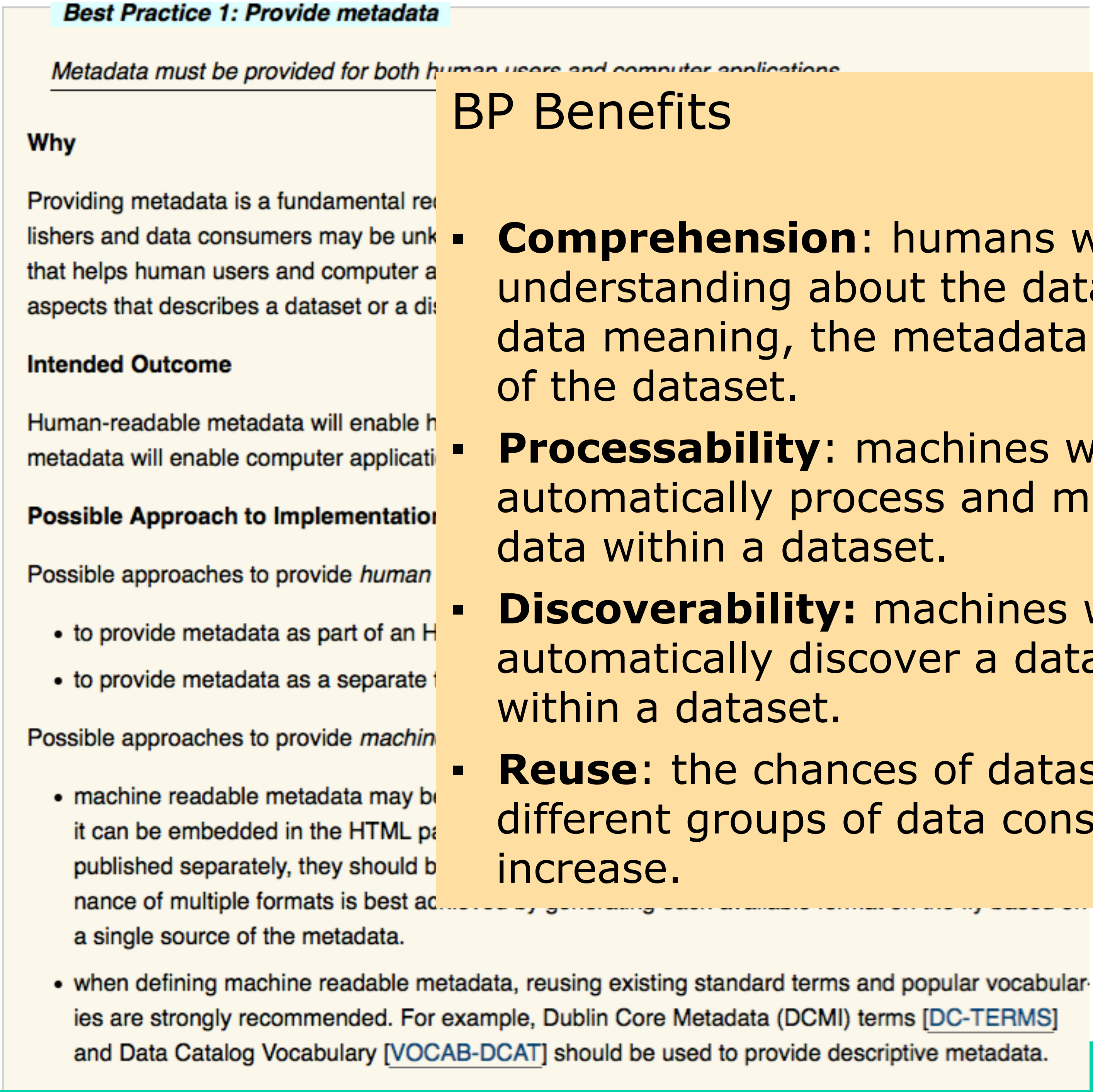
BP: Provide metadata
BP: Provide locale parameters metadata
BP: Provide structural metadata
BP: Provide descriptive metadata

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

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 re
lishers and data consumers may be unk
that helps human users and computer a
aspects that describes a dataset or a di

Intended Outcome

Human-readable metadata will enable h
metadata will enable computer applicati

Possible Approach to Implementatio

Possible approaches to provide *human*

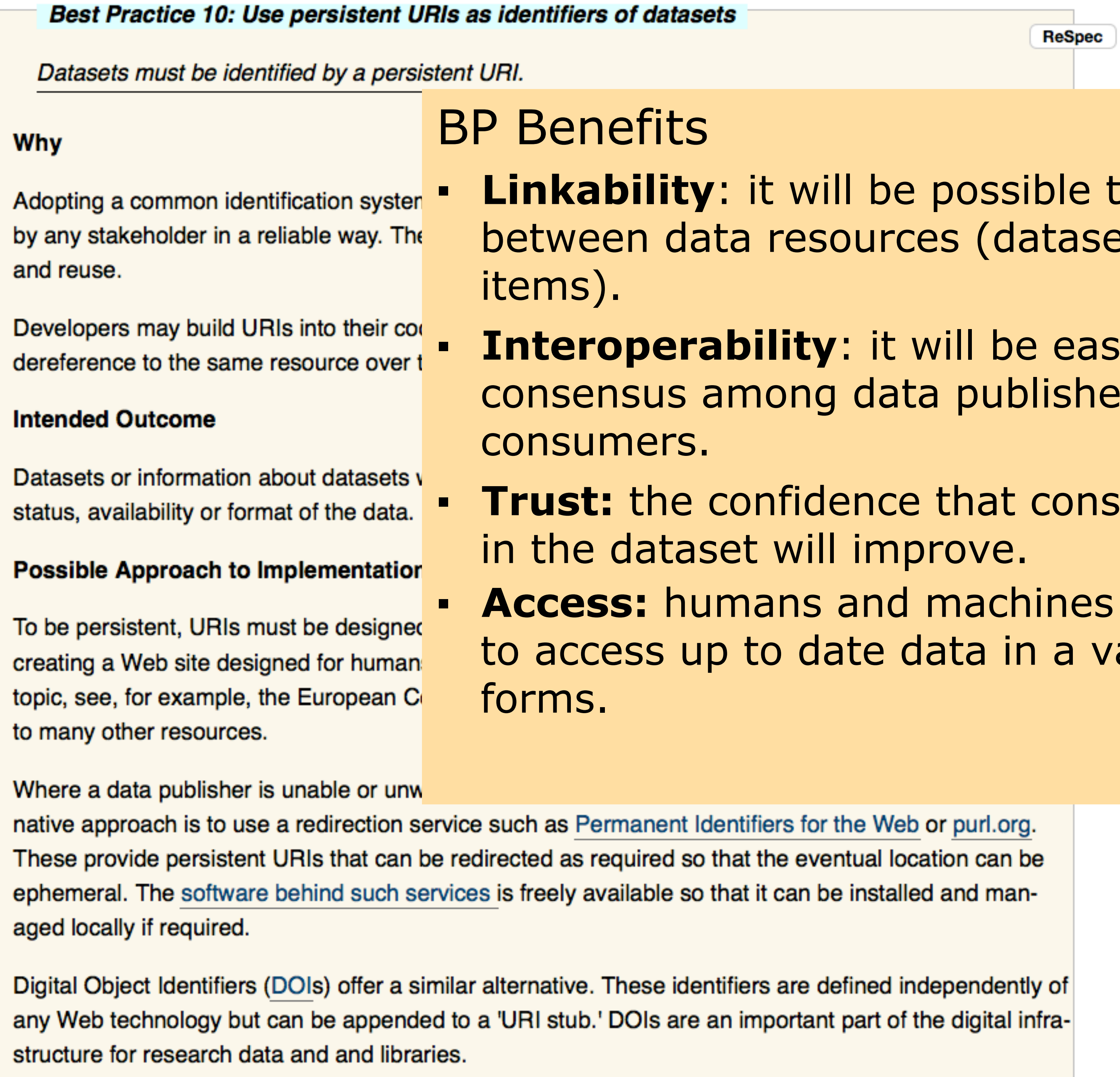
- to provide metadata as part of an H
- to provide metadata as a separate

Possible approaches to provide *machin*

- machine readable metadata may b
it can be embedded in the HTML pa
published separately, they should b
nance of multiple formats is best ac
a single source of the metadata.
- when defining machine readable metadata, reusing existing standard terms and popular vocabularies are strongly recommended. For example, Dublin Core Metadata (DCMI) terms [DC-TERMS] and Data Catalog Vocabulary [VOCAB-DCAT] should be used to provide descriptive metadata.

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.



Best Practice 10: Use persistent URIs as identifiers of datasets

Datasets must be identified by a persistent URI.

Why

Adopting a common identification system

by any stakeholder in a reliable way. The

and reuse.

Developers may build URIs into their co

dereference to the same resource over t

Intended Outcome

Datasets or information about datasets v

status, availability or format of the data.

Possible Approach to Implementation

To be persistent, URIs must be designed

creating a Web site designed for human

topic, see, for example, the European C

to many other resources.

Where a data publisher is unable or unw

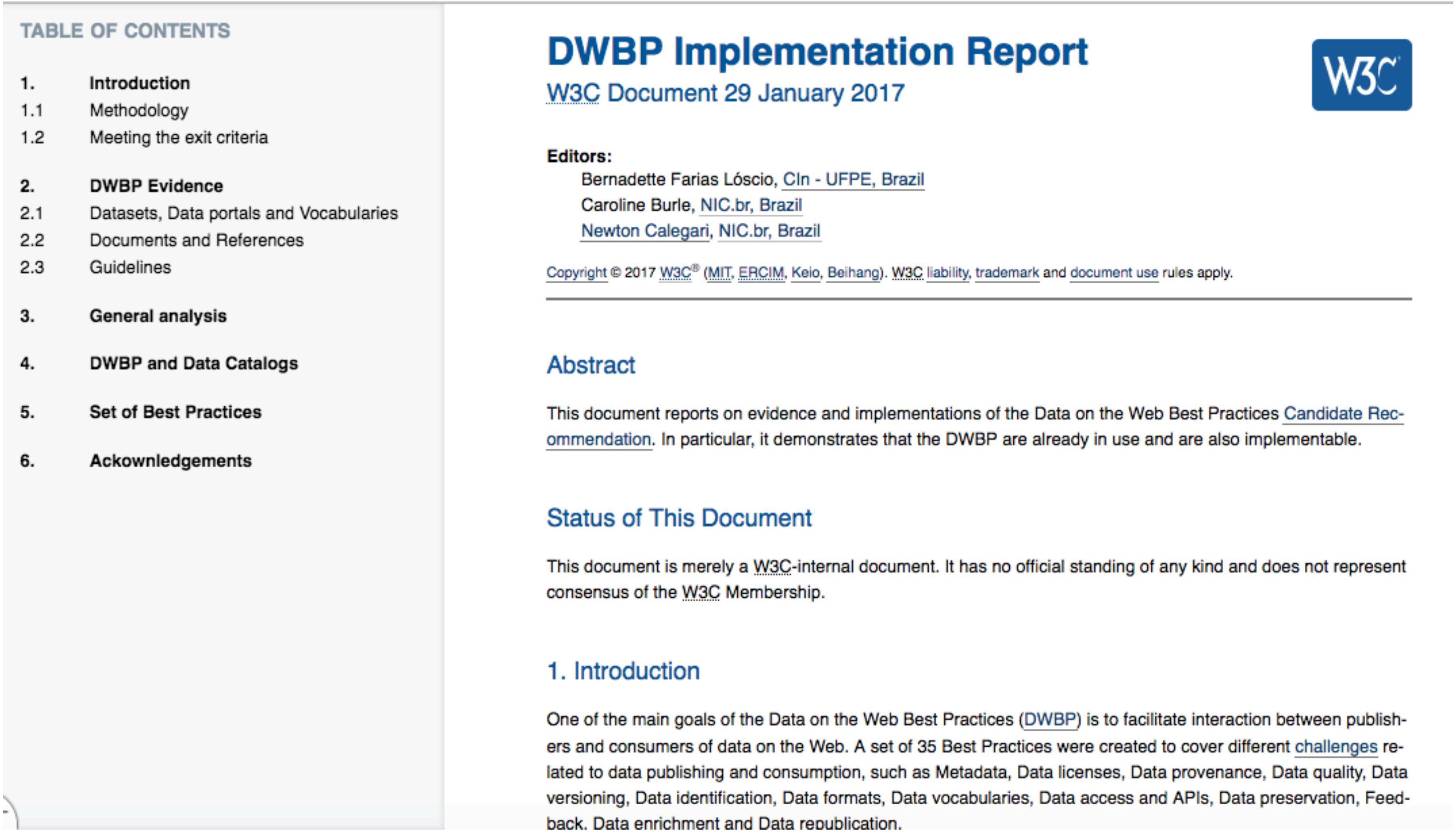
native approach is to use a redirection service such as Permanent Identifiers for the Web or purl.org. These provide persistent URIs that can be redirected as required so that the eventual location can be ephemeral. The software behind such services is freely available so that it can be installed and managed locally if required.

Digital Object Identifiers (DOIs) offer a similar alternative. These identifiers are defined independently of any Web technology but can be appended to a 'URI stub.' DOIs are an important part of the digital infrastructure for research data and and libraries.

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?



DWBP Implementation Report

W3C Document 29 January 2017

Editors:
Bernadette Farias Lóscio, CIn - UFPE, Brazil
Caroline Burle, 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

Obrigada(o)!

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

cburle@nic.br
bfl@cin.ufpe.br
newton@nic.br

@carolburle
@bernafarias
@newtoncalegari