Publishing data on the Web

Bernadette Farias Lóscio and Caroline Burle

Centro de Informática U·F·P·E

ceweb.br
Centro de Estudos sobre Tecnologias Web

nic.br

nic.br cgi.br

Núcleo de Informação
e Coordenação do
Ponto BR

Comitê Gestor da
Internet no Brasil

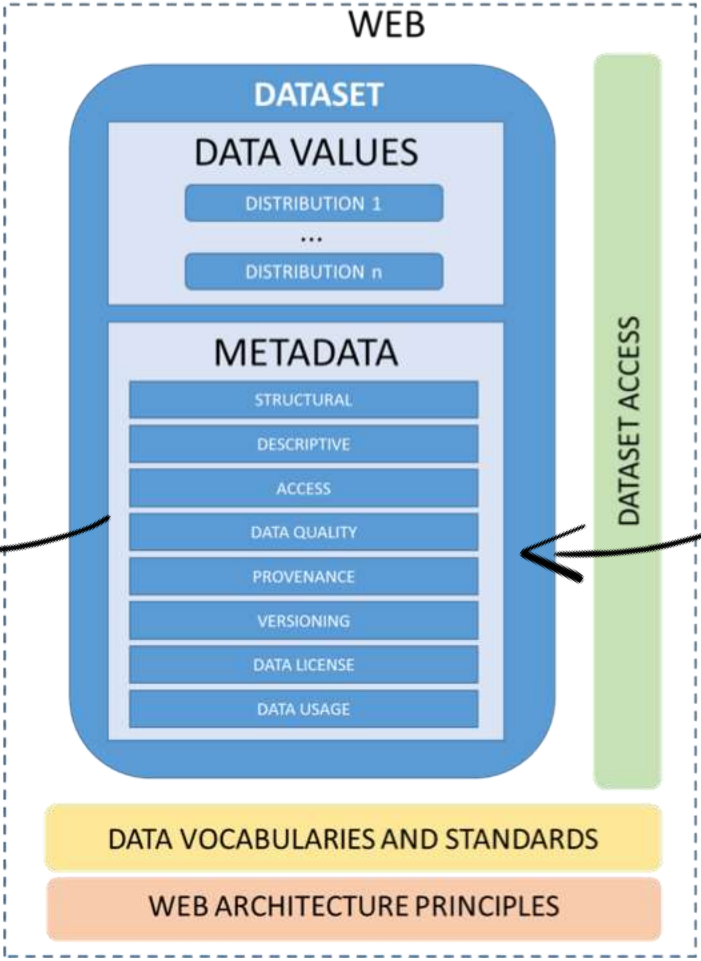
registro.br cert.br cetic.br ceptro.br ptt.br ceweb.br

Topics to be discussed

- Data on the Web
- Linked Data
- Open Data
- Big Data, Open Data and Data on the Web
- Data on the Web lifecycle
- DWBP: Challenges and Benefits
- Questions and comments

Data on the Web

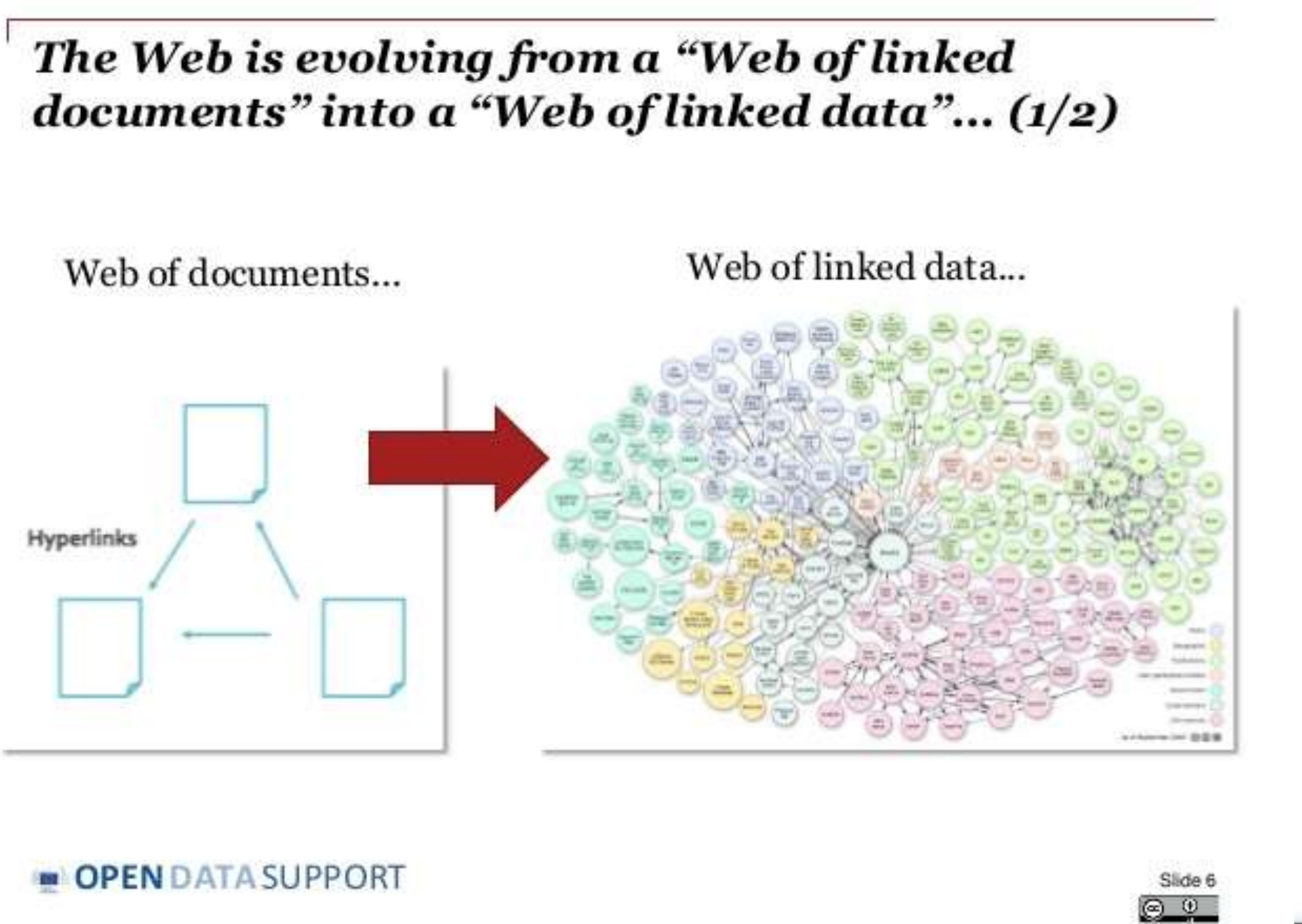
Consumes data



Publishing data on the web

Publishes data

Web of Linked Documents → Web of Linked Data



The Web is evolving from a "Web of linked documents" into a "Web of linked data"... (1/2)

Web of documents...

Web of linked data...

OPEN DATA SUPPORT

Slide 6

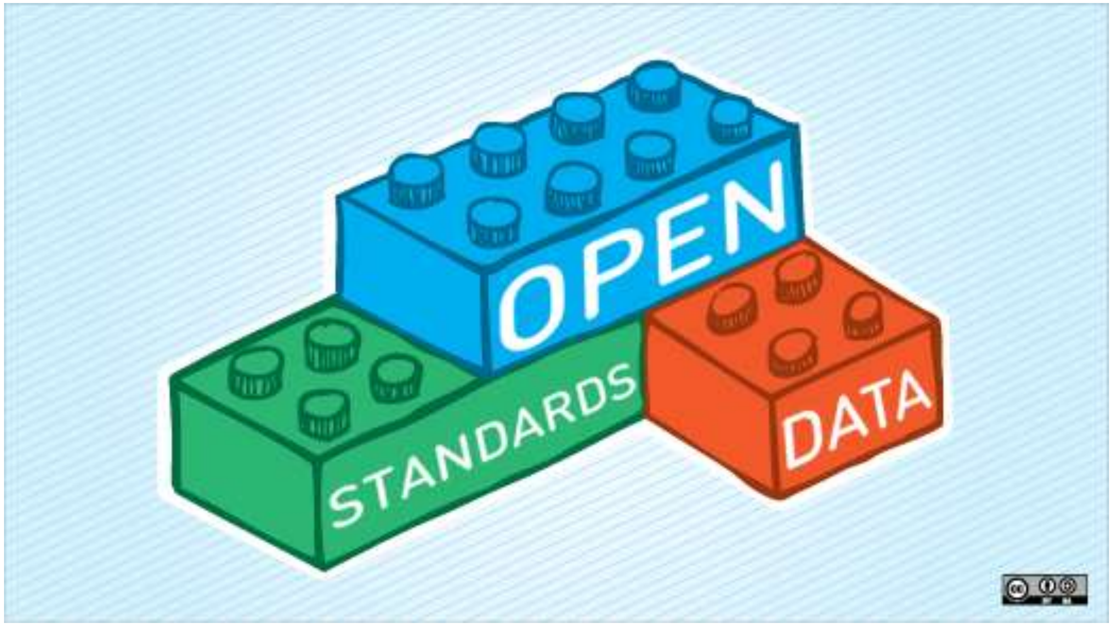
What is Open Data?

Characteristics:

Availability and access

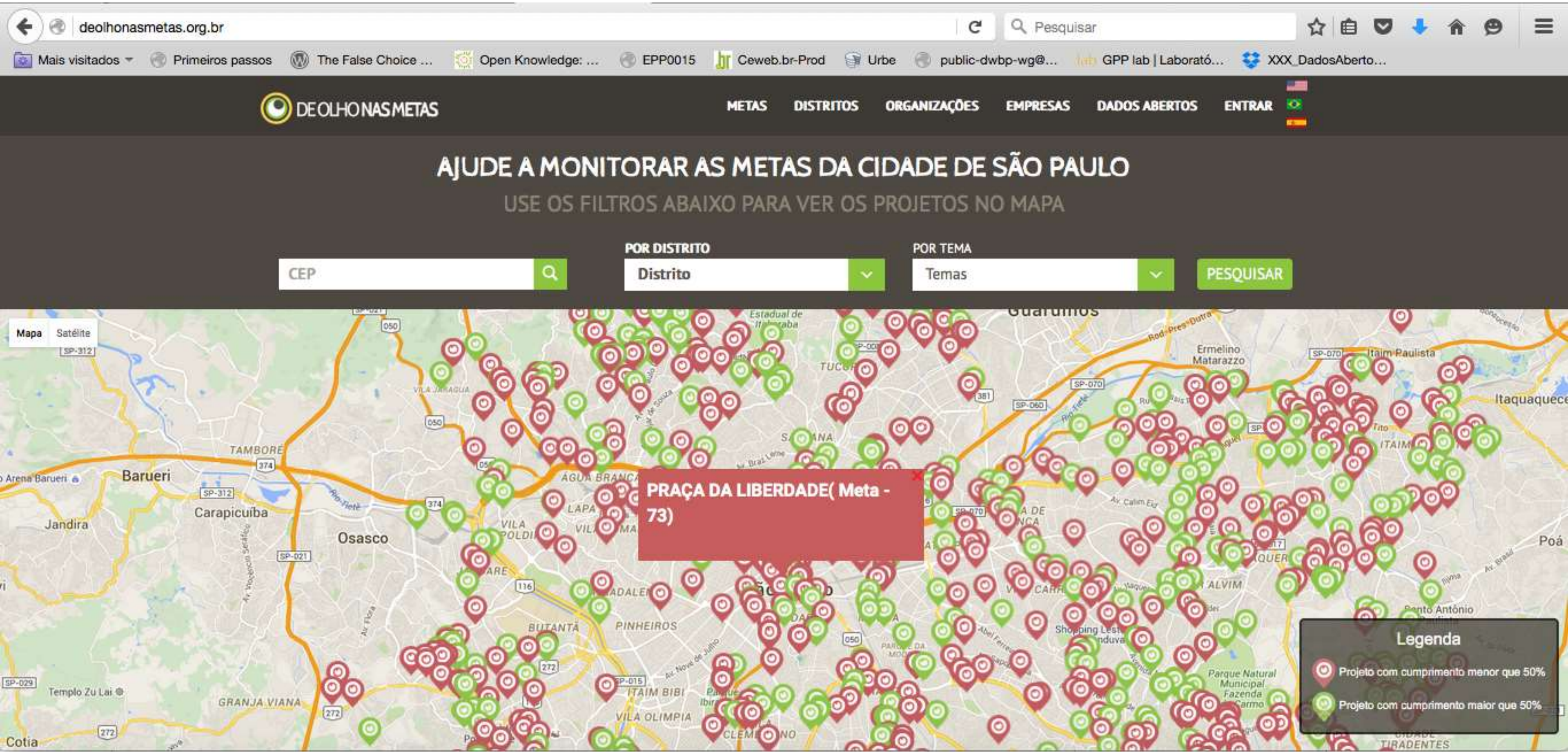
Reuse and Redistribution

Universal Participation



Source: http://ceweb.br/publicacao/open-data-guideline/

De Olho nas Metas



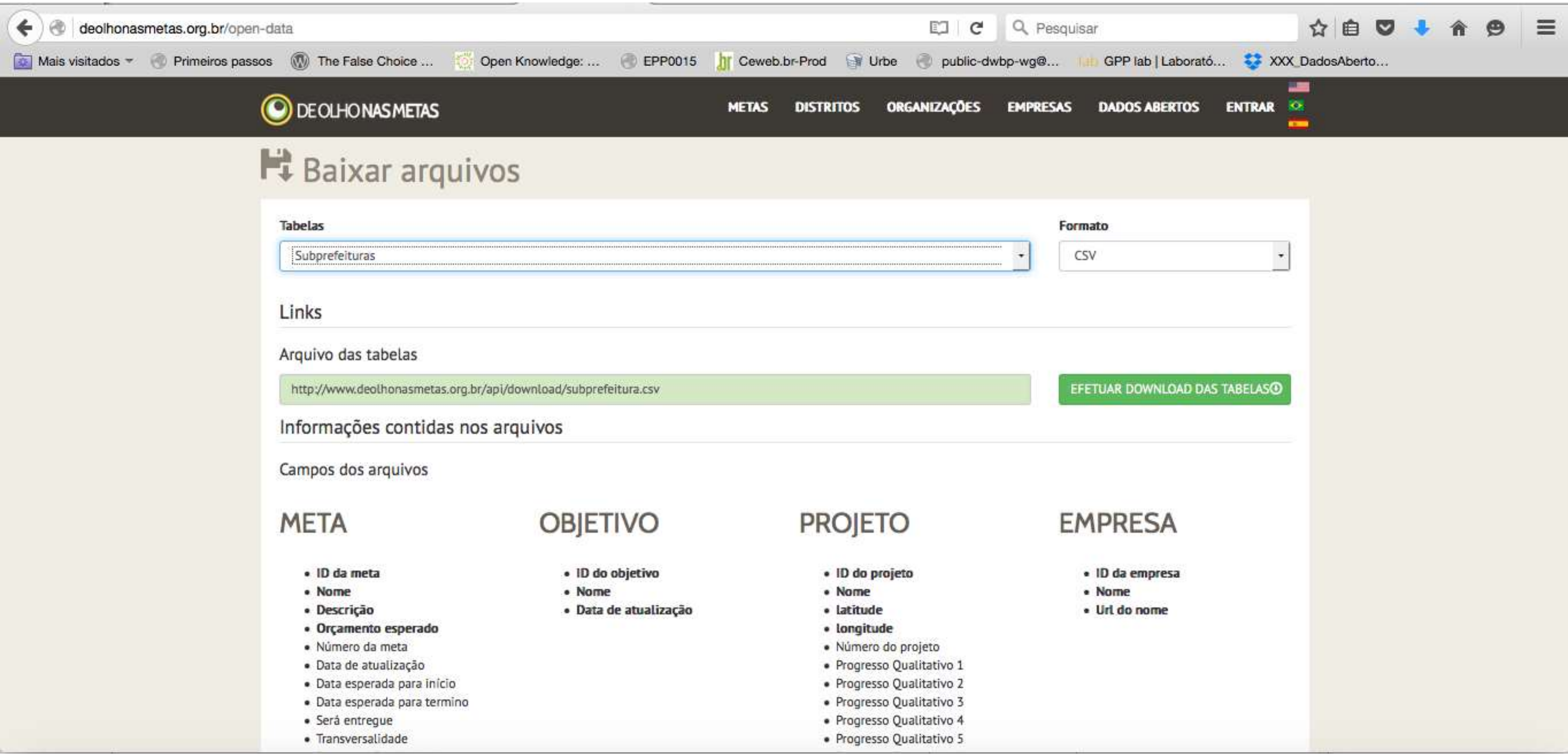
AJUDE A MONITORAR AS METAS DA CIDADE DE SÃO PAULO

USE OS FILTROS ABAIXO PARA VER OS PROJETOS NO MAPA

PRAÇA DA LIBERDADE(Meta - 73)

Source: http://deolhonasmetas.org.br

De Olho nas Metas



Baixar arquivos

Tabelas: Subprefeituras

Formato: CSV

Links

Arquivo das tabelas

http://www.deolhonasmetas.org.br/api/download/subprefeitura.csv

EFETUAR DOWNLOAD DAS TABELAS

Informações contidas nos arquivos

Campos dos arquivos

META

- **ID da meta**
- **Nome**
- **Descrição**
- **Orçamento esperado**
- Número da meta
- Data de atualização
- Data esperada para início
- Data esperada para termino
- Será entregue
- Transversalidade

OBJETIVO

- **ID do objetivo**
- **Nome**
- **Data de atualização**

PROJETO

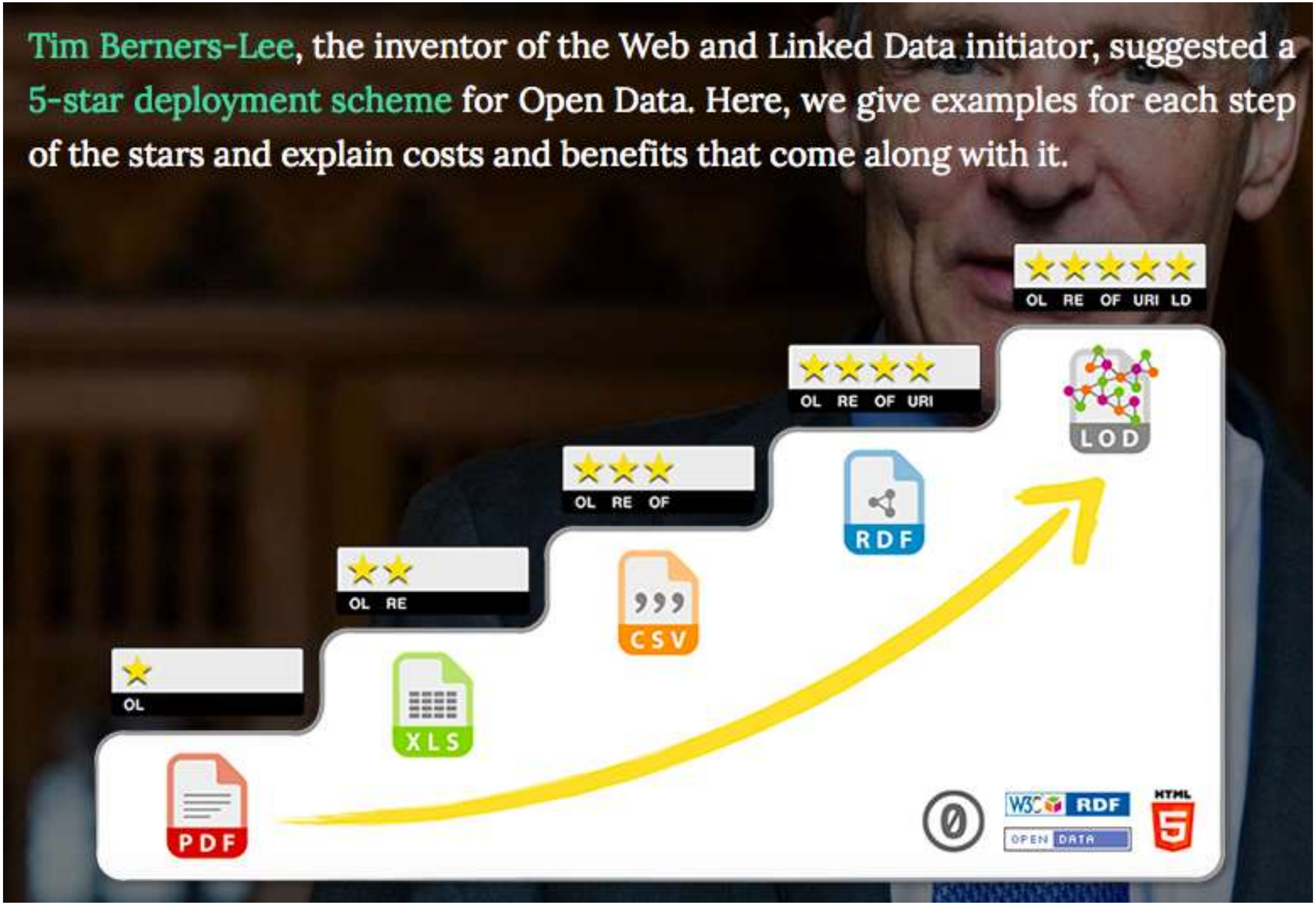
- **ID do projeto**
- **Nome**
- **latitude**
- **longitude**
- Número do projeto
- Progresso Qualitativo 1
- Progresso Qualitativo 2
- Progresso Qualitativo 3
- Progresso Qualitativo 4
- Progresso Qualitativo 5

EMPRESA

- **ID da empresa**
- **Nome**
- **Url do nome**

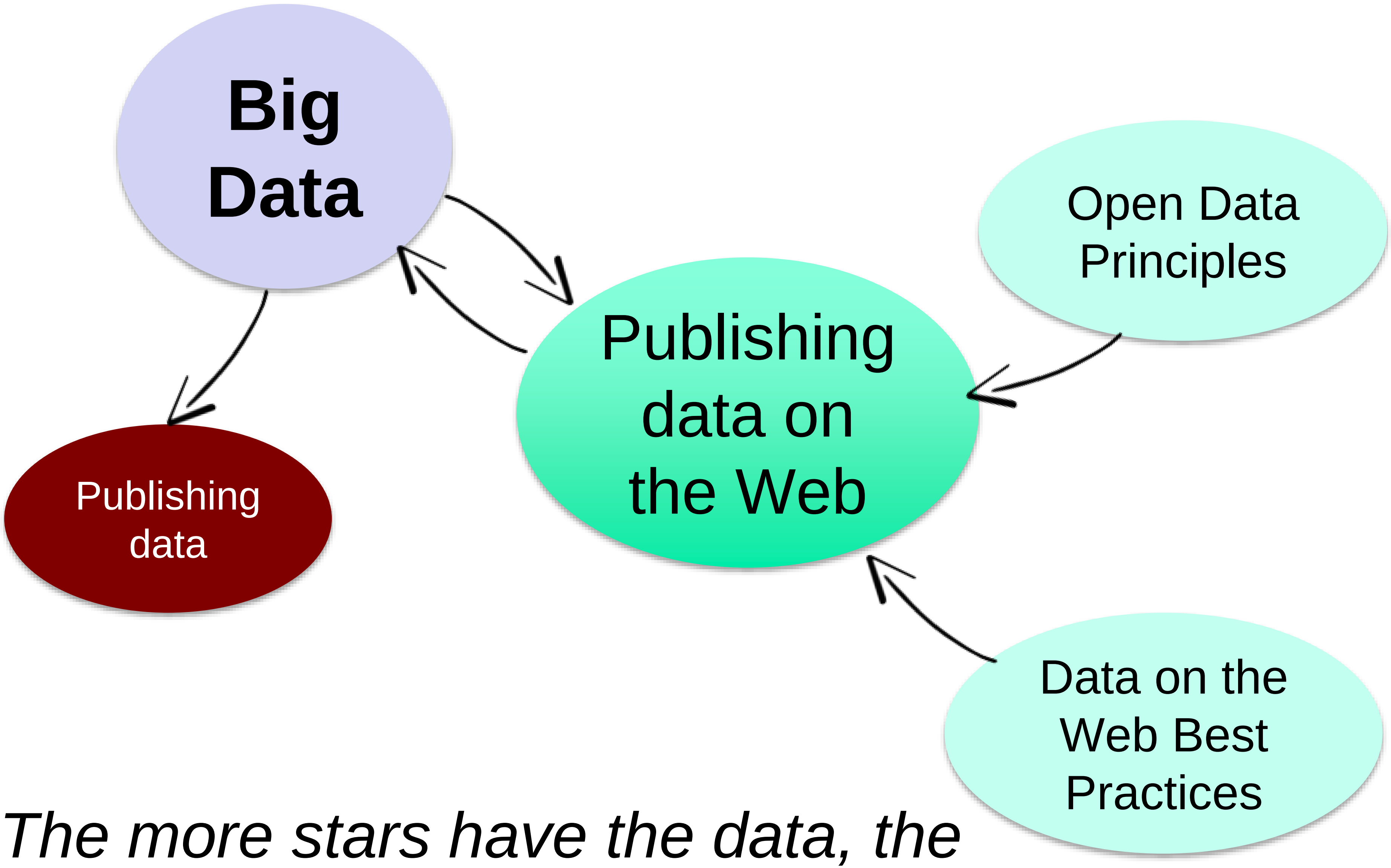
Source: http://deolhonasmetas.org.br/open-data

5-Star Open Data



Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.

Source: http://5stardata.info/



Big Data

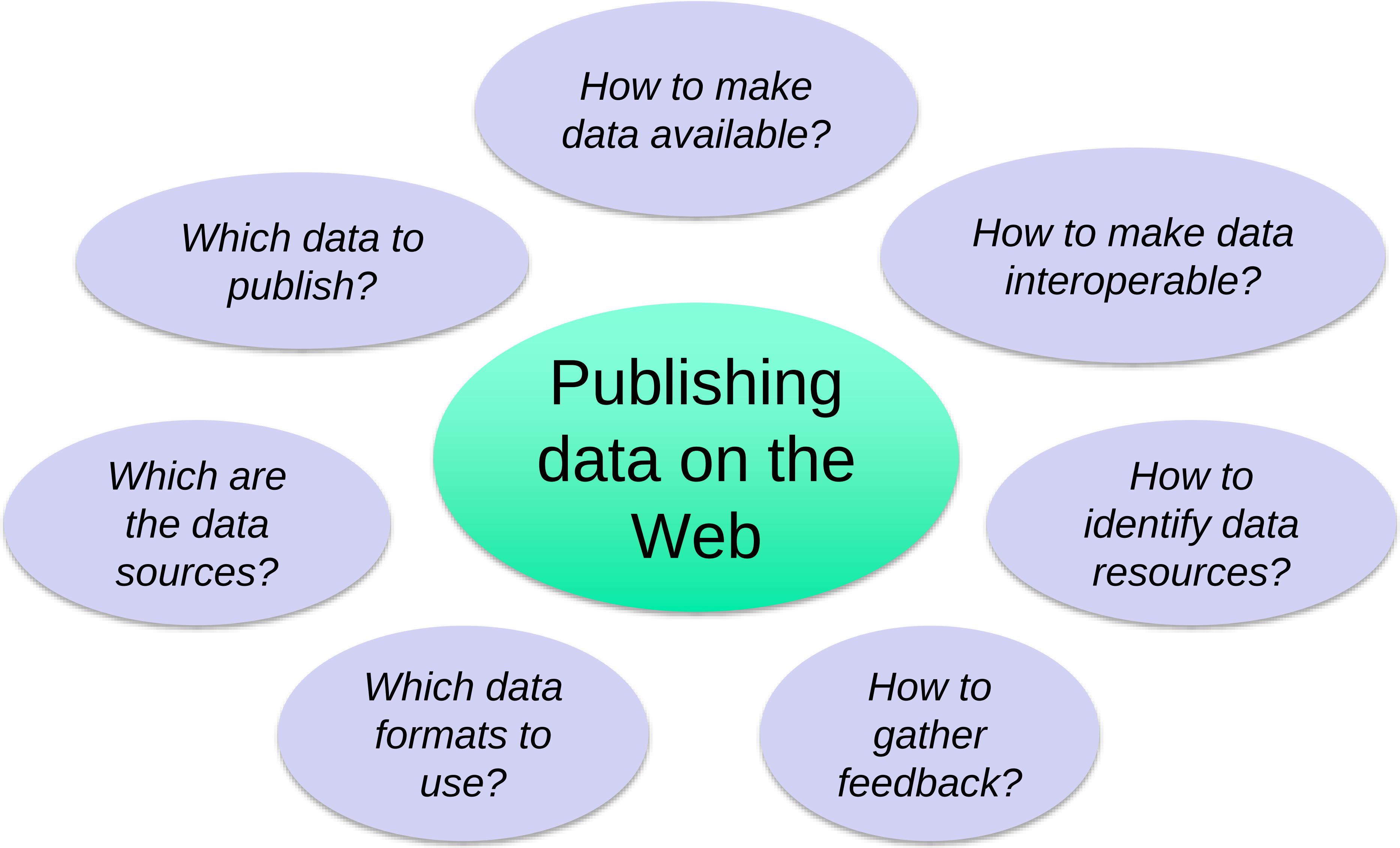
Publishing data

Publishing data on the Web

Open Data Principles

Data on the Web Best Practices

The more stars have the data, the easier it is to find and reuse the data.



Publishing data on the Web

How to make data available?

How to make data interoperable?

How to identify data resources?

How to gather feedback?

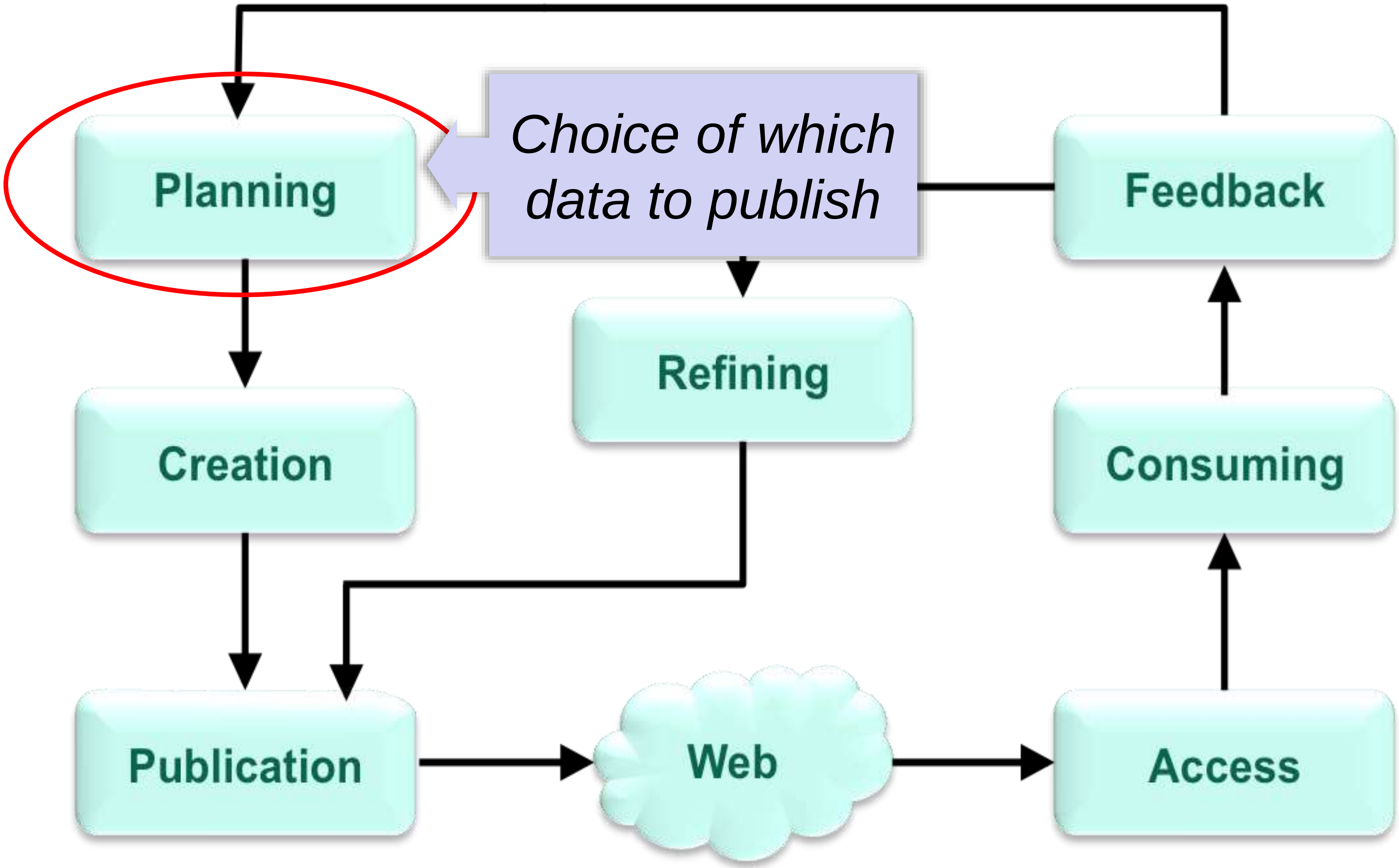
Which data formats to use?

Which are the data sources?

Which data to publish?

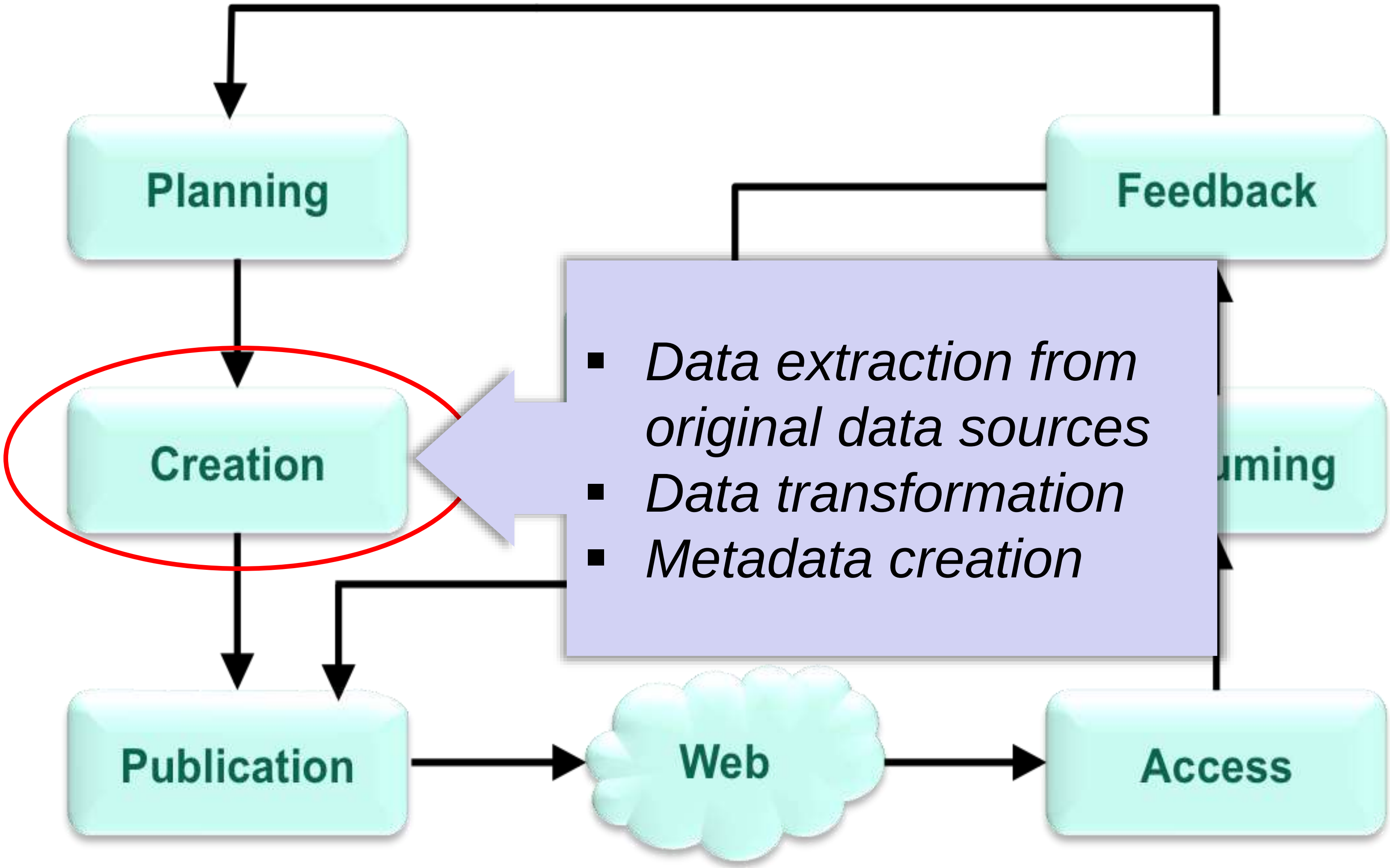
Publishing data on the Web is more than just publishing!

Data on the Web Lifecycle



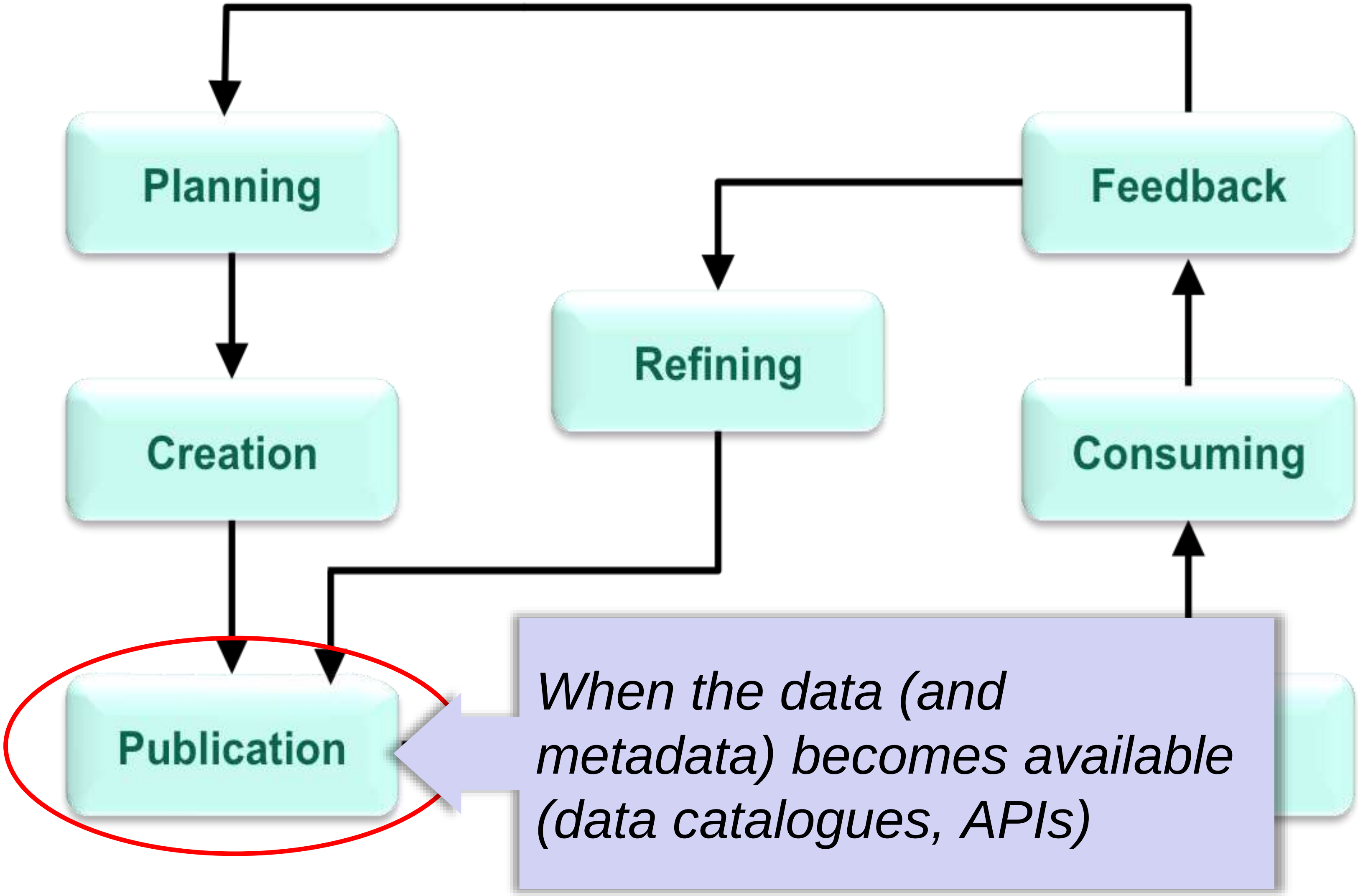
Choice of which data to publish

Data on the Web Lifecycle



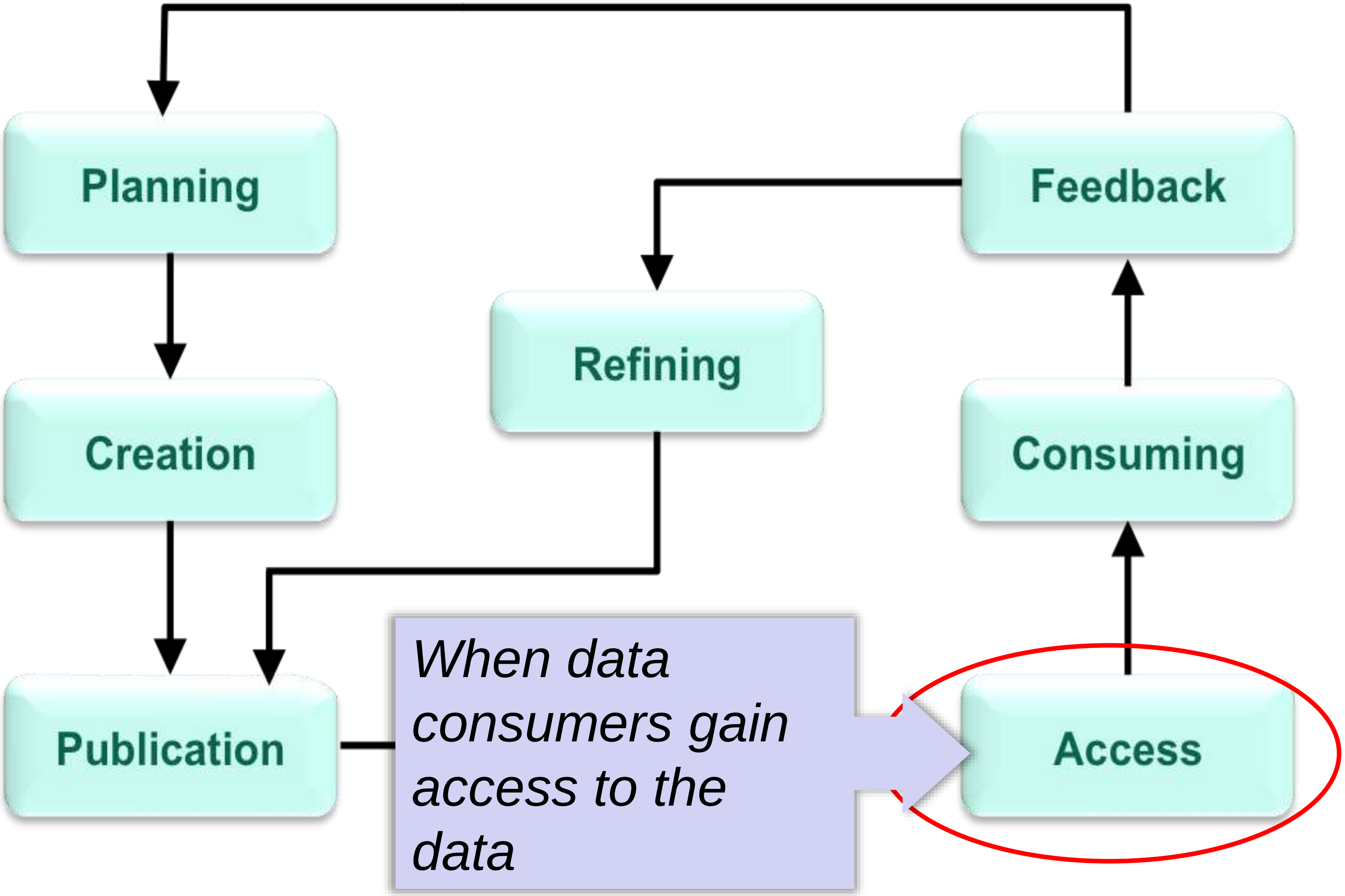
- *Data extraction from original data sources*
- *Data transformation*
- *Metadata creation*

Data on the Web Lifecycle



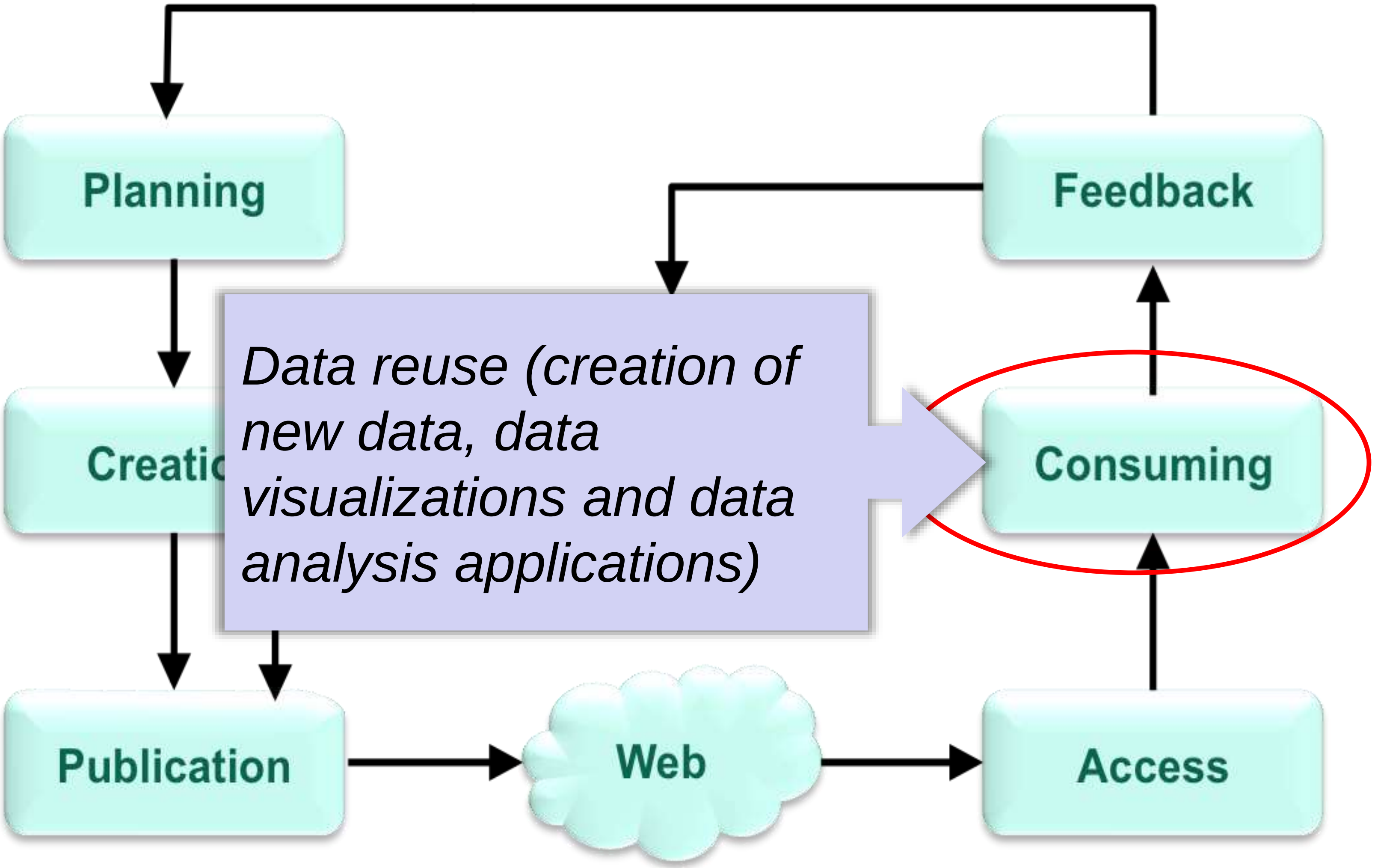
When the data (and metadata) becomes available (data catalogues, APIs)

Data on the Web Lifecycle



When data consumers gain access to the data

Data on the Web Lifecycle



Planning

Creation

Publication

Web

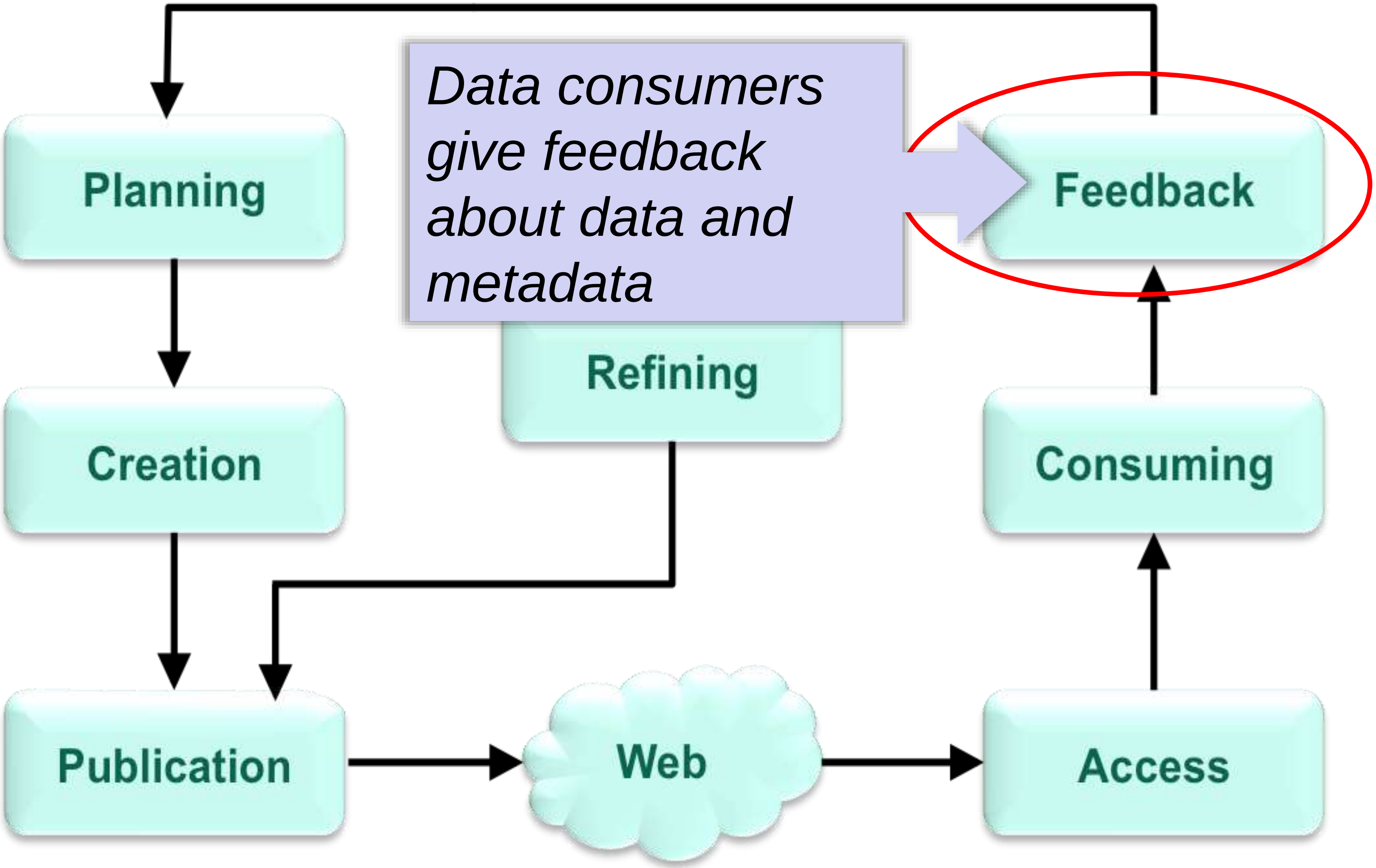
Access

Consuming

Feedback

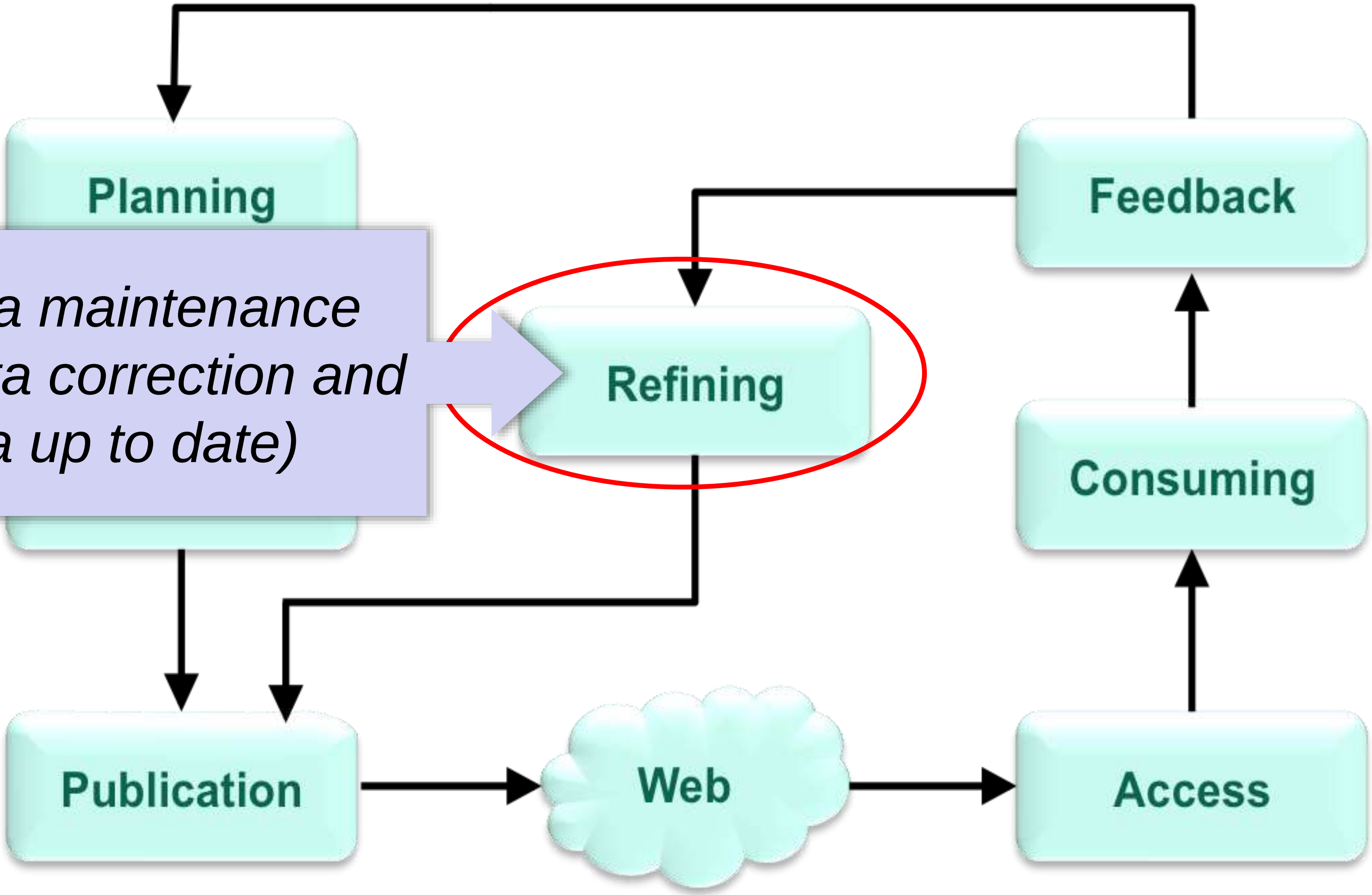
Data reuse (creation of new data, data visualizations and data analysis applications)

Data on the Web Lifecycle



Data consumers give feedback about data and metadata

Data on the Web Lifecycle



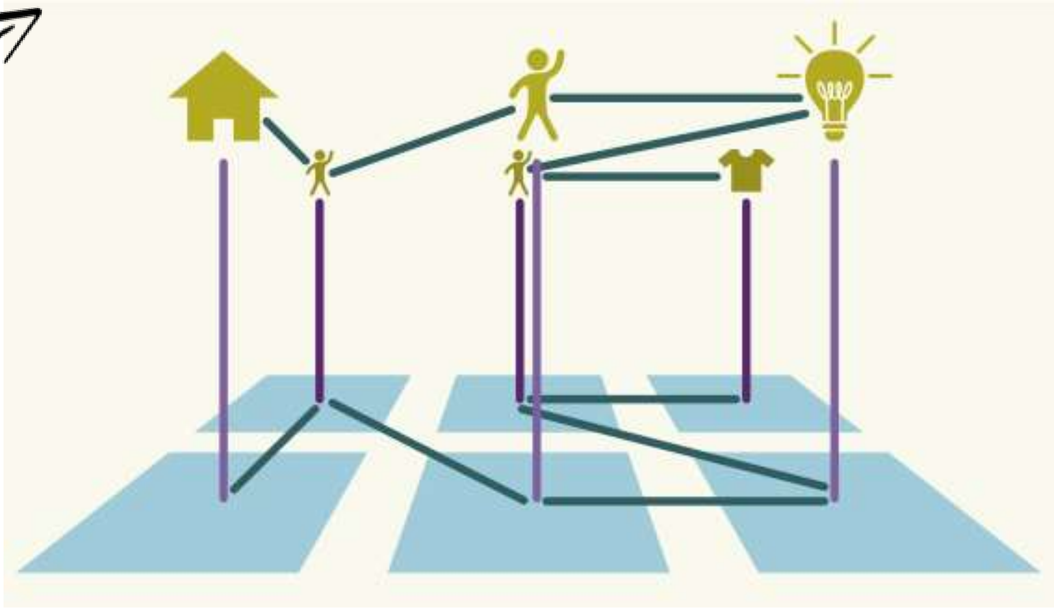
Data maintenance (data correction and data up to date)

Players of the data on the Web ecosystem

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

Data publisher:
publishes and shares data

Data consumer:
reuses the data and might generate new data



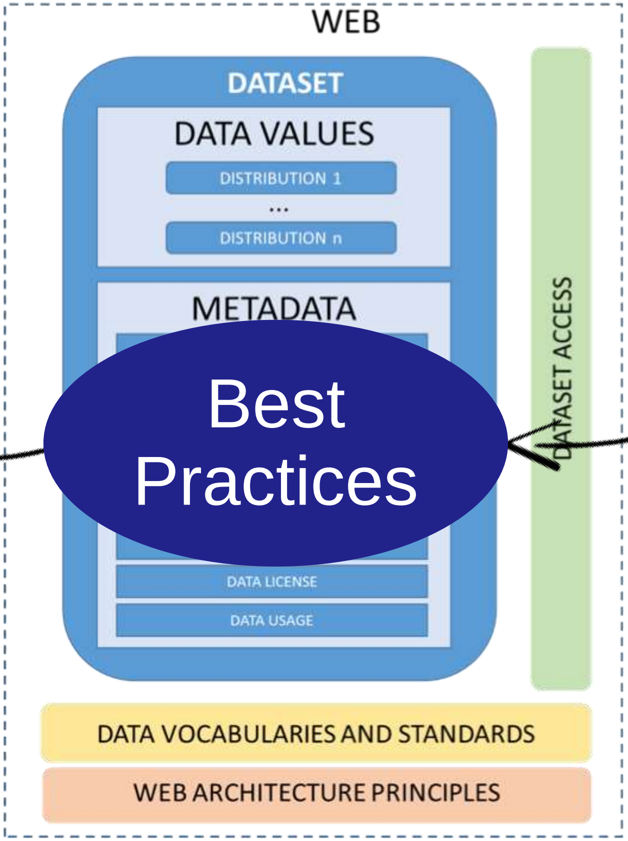
Source: http://ceweb.br/livros/dados-abertos-conectados/capitulo-1/

How to enable the data reuse?

How to enable the reuse of data?

A common understanding between data publishers and data consumers becomes fundamental.
Without this agreement, data publishers' efforts may be incompatible with data consumers' desires.

Consumes data



WEB

DATASET

DATA VALUES

DISTRIBUTION 1

...

DISTRIBUTION n

METADATA

Best Practices

DATA LICENSE

DATA USAGE

DATASET ACCESS

DATA VOCABULARIES AND STANDARDS

WEB ARCHITECTURE PRINCIPLES

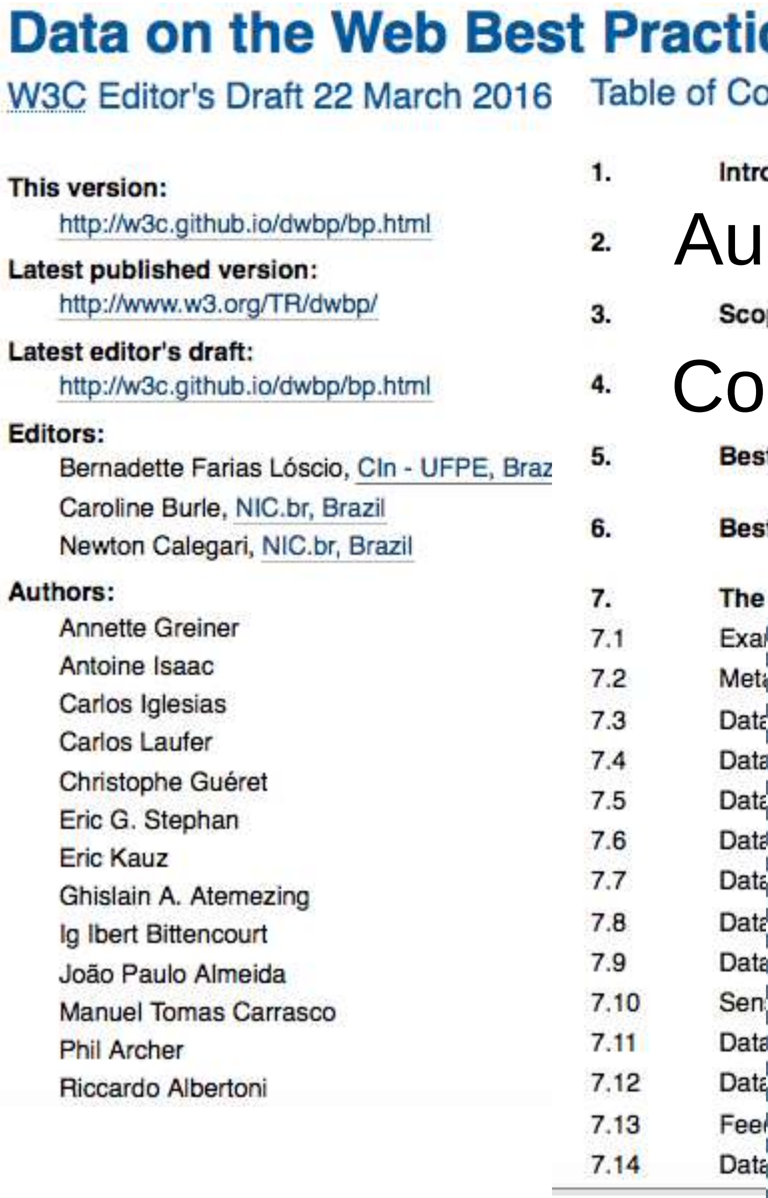
Publishes data

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 Best Practices

W3C Editor's Draft 22 March 2016

This version:
http://w3c.github.io/dwbp/bp.html

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

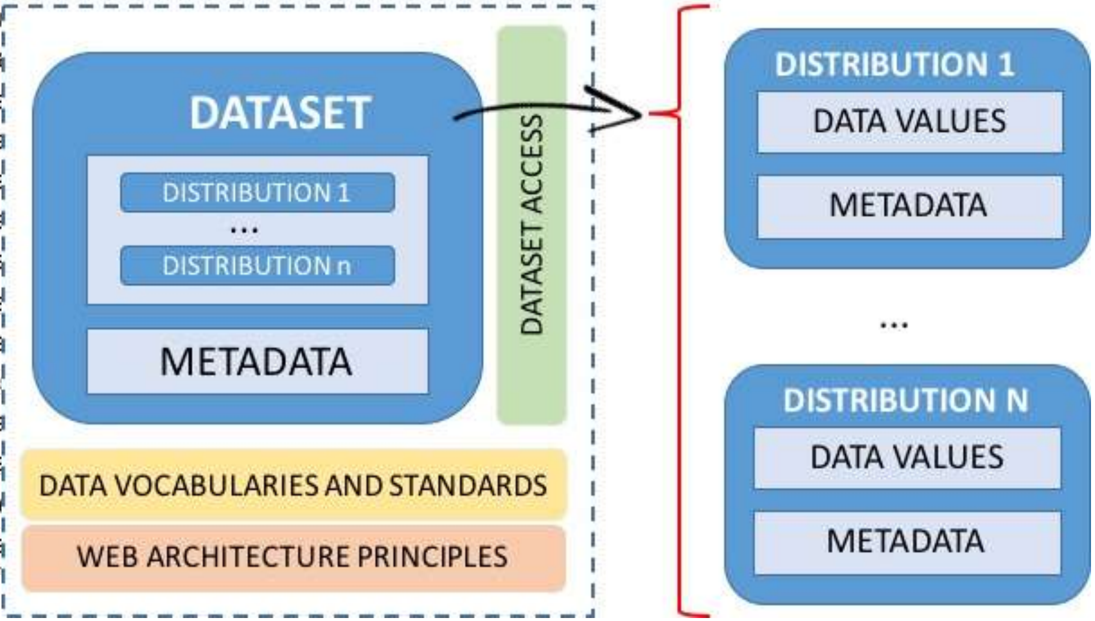
Latest editor's draft:
http://w3c.github.io/dwbp/bp.html

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

Authors:
Annette Greiner
Antoine Isaac
Carlos Iglesias
Carlos Laufer
Christophe Guéret
Eric G. Stephan
Eric Kauz
Ghislain A. Atemezing
Ig Ibert Bittencourt
João Paulo Almeida
Manuel Tomas Carrasco
Phil Archer
Riccardo Albertoni

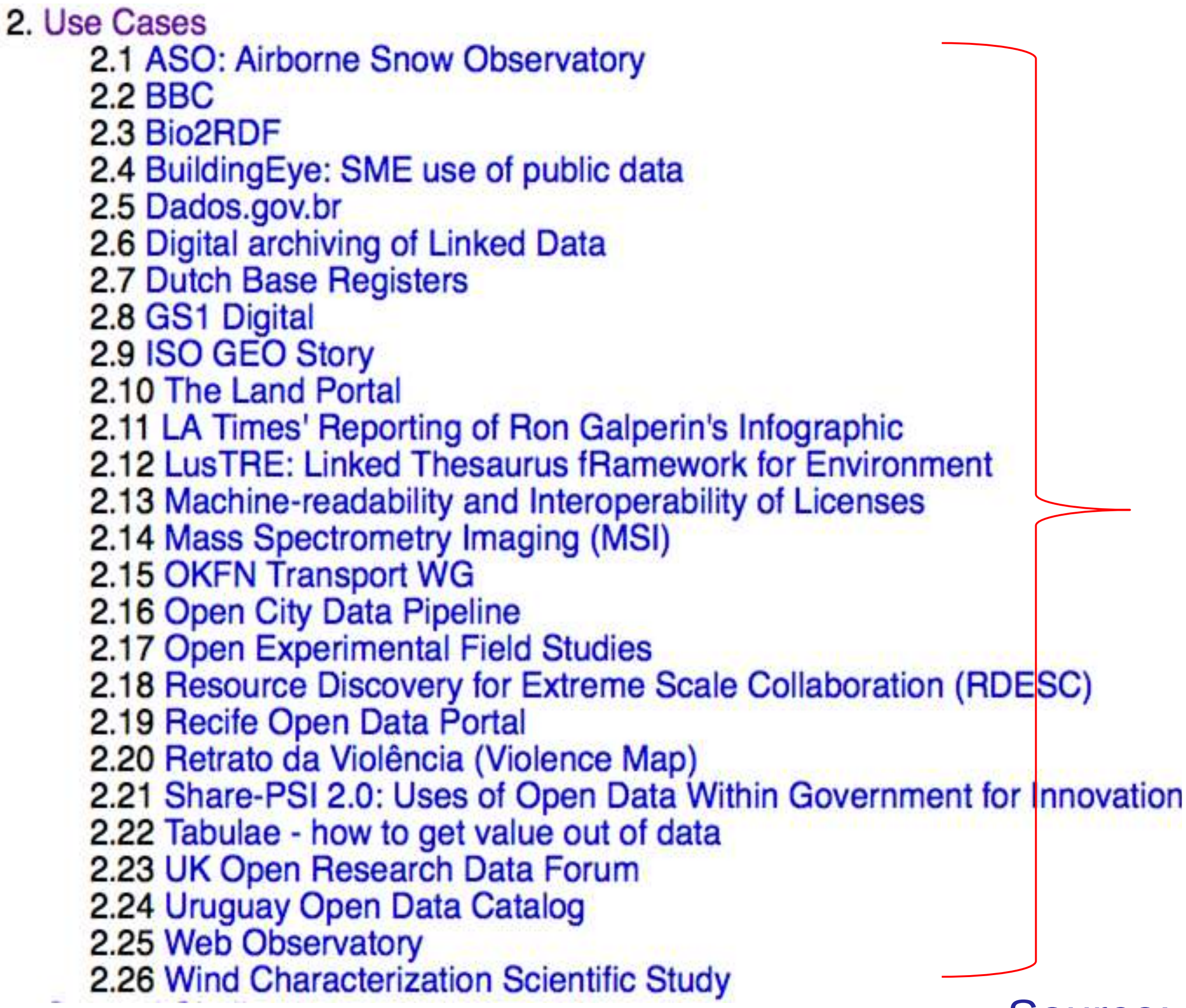
Table of Contents

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

Data on the Web Use cases & Requirements



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
   - 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

scenarios of how data is commonly published on the Web and how it is used

cover different domains and illustrate some of the main challenges faced by data publishers and data consumers

Source: https://www.w3.org/TR/dwbp-ucr/

Data on the Web Challenges

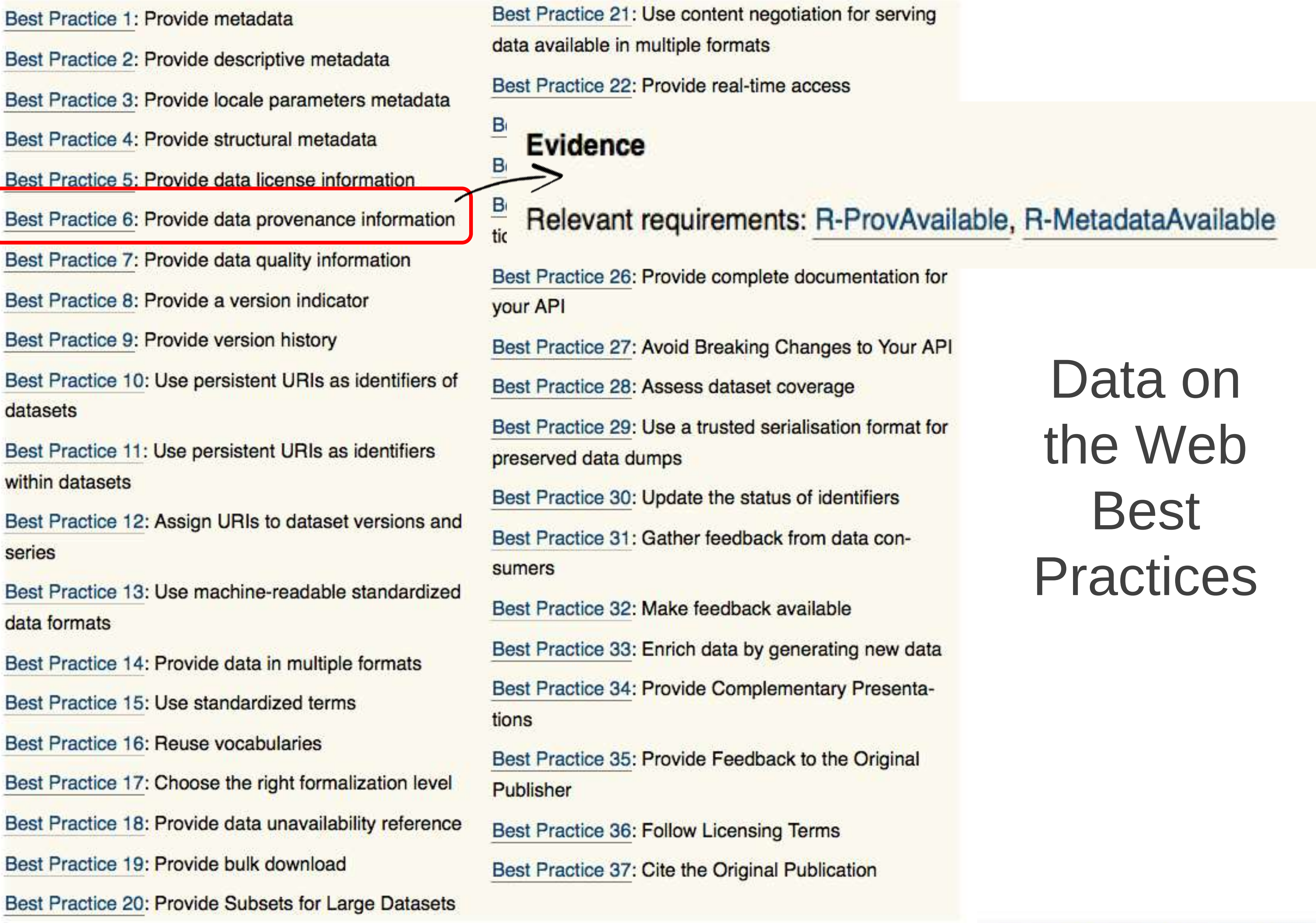
- Metadata *(for humans & machines)*
- Data Licenses *(how to permite & restrict access?)*
- Data Provenance & Quality *(how to add trust?)*
- Data Versioning *(tracking dataset versions)*
- Data Identifiers *(identifying datasets and distributions)*
- Data Formats *(which data formats to use?)*

The openness and flexibility of the Web creates new challenges for data publishers and data consumers

Data on the Web Challenges

- Data Vocabularies *(how to promote interoperability?)*
- Sensitive Data *(Privacy & Security)*
- Data Access *(access options)*
- Feedback *(how to engage users?)*
- Data Enrichment *(adding value to data)*

Data on the Web Best Practices



Best Practice 1: Provide metadata

Best Practice 2: Provide descriptive metadata

Best Practice 3: Provide locale parameters metadata

Best Practice 4: Provide structural metadata

Best Practice 5: Provide data license information

Best Practice 6: Provide data provenance information

Best Practice 7: Provide data quality information

Best Practice 8: Provide a version indicator

Best Practice 9: Provide version history

Best Practice 10: Use persistent URIs as identifiers of datasets

Best Practice 11: Use persistent URIs as identifiers within datasets

Best Practice 12: Assign URIs to dataset versions and series

Best Practice 13: Use machine-readable standardized data formats

Best Practice 14: Provide data in multiple formats

Best Practice 15: Use standardized terms

Best Practice 16: Reuse vocabularies

Best Practice 17: Choose the right formalization level

Best Practice 18: Provide data unavailability reference

Best Practice 19: Provide bulk download

Best Practice 20: Provide Subsets for Large Datasets

Best Practice 21: Use content negotiation for serving data available in multiple formats

Best Practice 22: Provide real-time access

B

B

B

tic

Evidence

Relevant requirements: R-ProvAvailable, R-MetadataAvailable

Best Practice 26: Provide complete documentation for your API

Best Practice 27: Avoid Breaking Changes to Your API

Best Practice 28: Assess dataset coverage

Best Practice 29: Use a trusted serialisation format for preserved data dumps

Best Practice 30: Update the status of identifiers

Best Practice 31: Gather feedback from data consumers

Best Practice 32: Make feedback available

Best Practice 33: Enrich data by generating new data

Best Practice 34: Provide Complementary Presentations

Best Practice 35: Provide Feedback to the Original Publisher

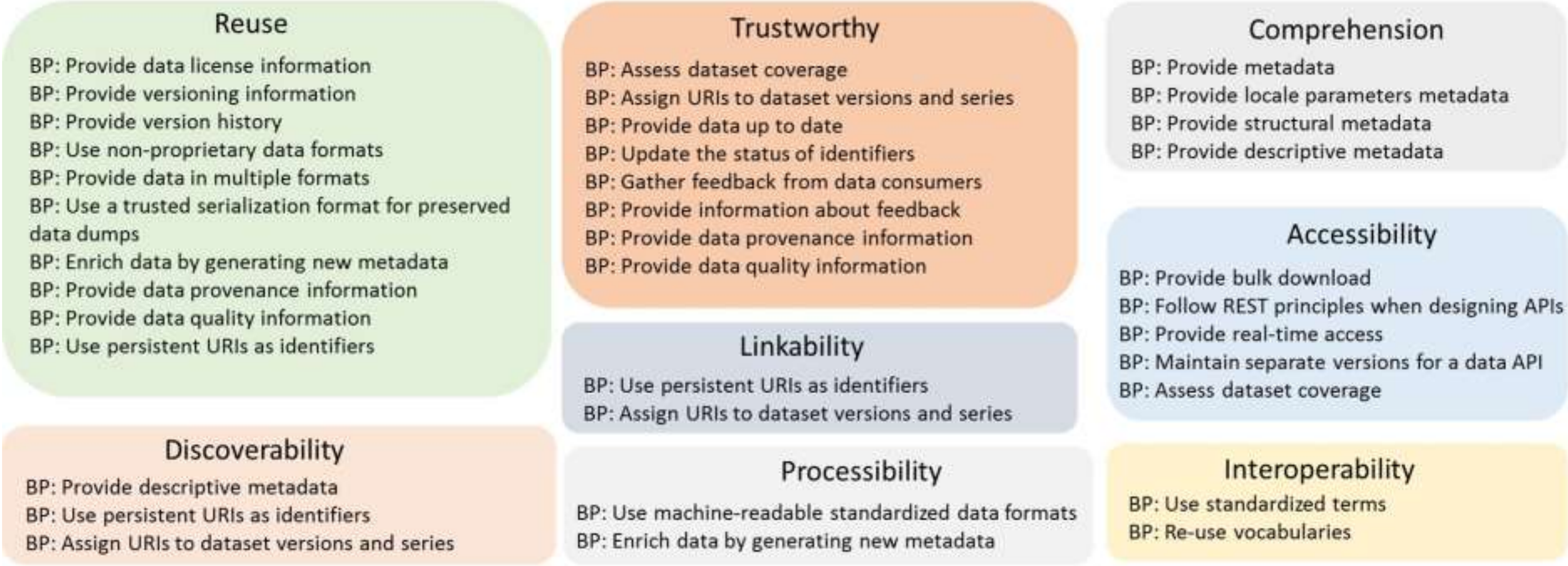
Best Practice 36: Follow Licensing Terms

Best Practice 37: Cite the Original Publication

DWBP Benefits

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

R
Reuse
C
Comprehension
L
Linkability
D
Discoverability
T
Trust
A
Access
I
Interoperability
P
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

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

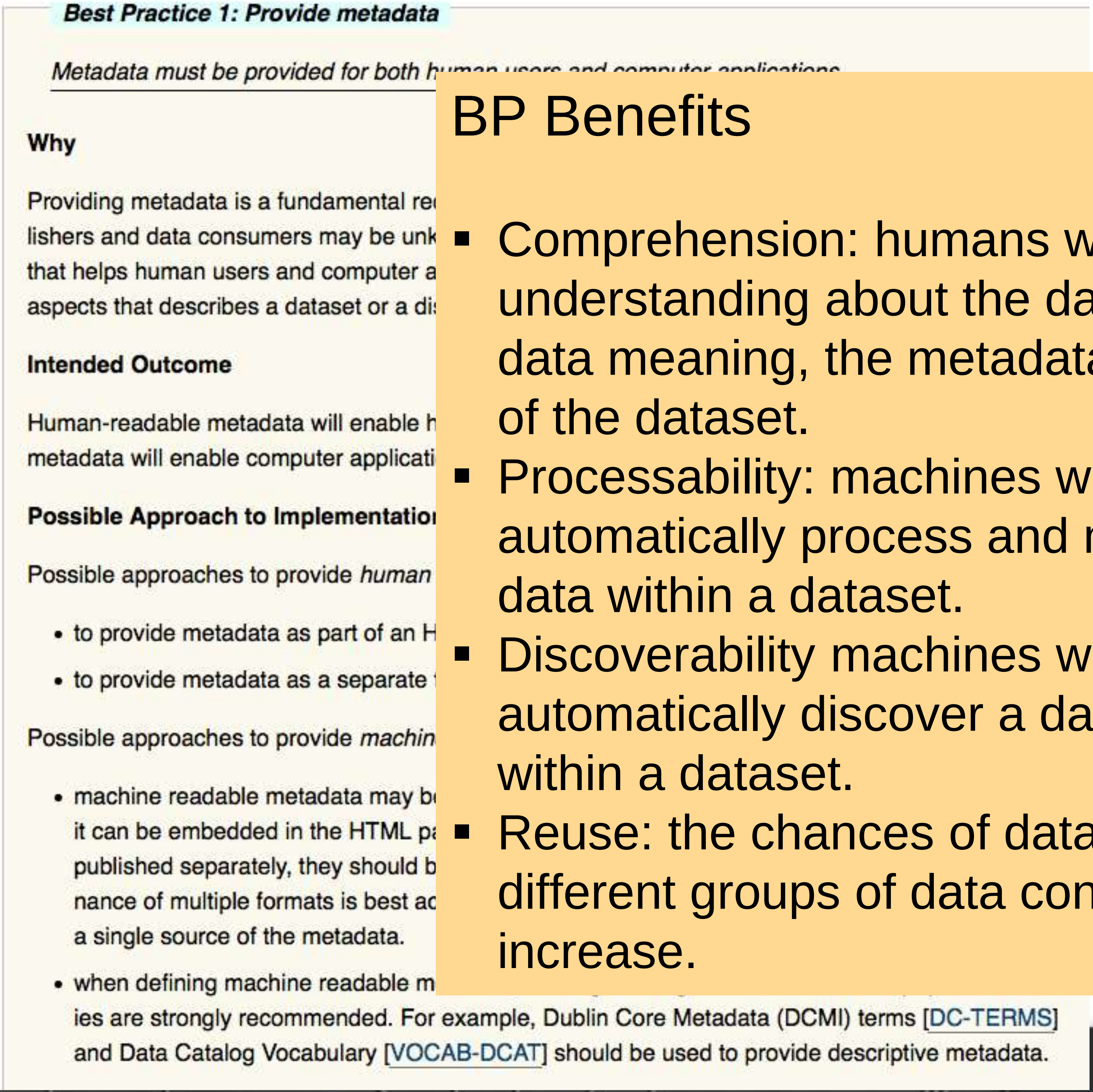
BP: Provide metadata
BP: Provide locale parameters metadata
BP: Provide structural metadata
BP: Provide descriptive metadata

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*

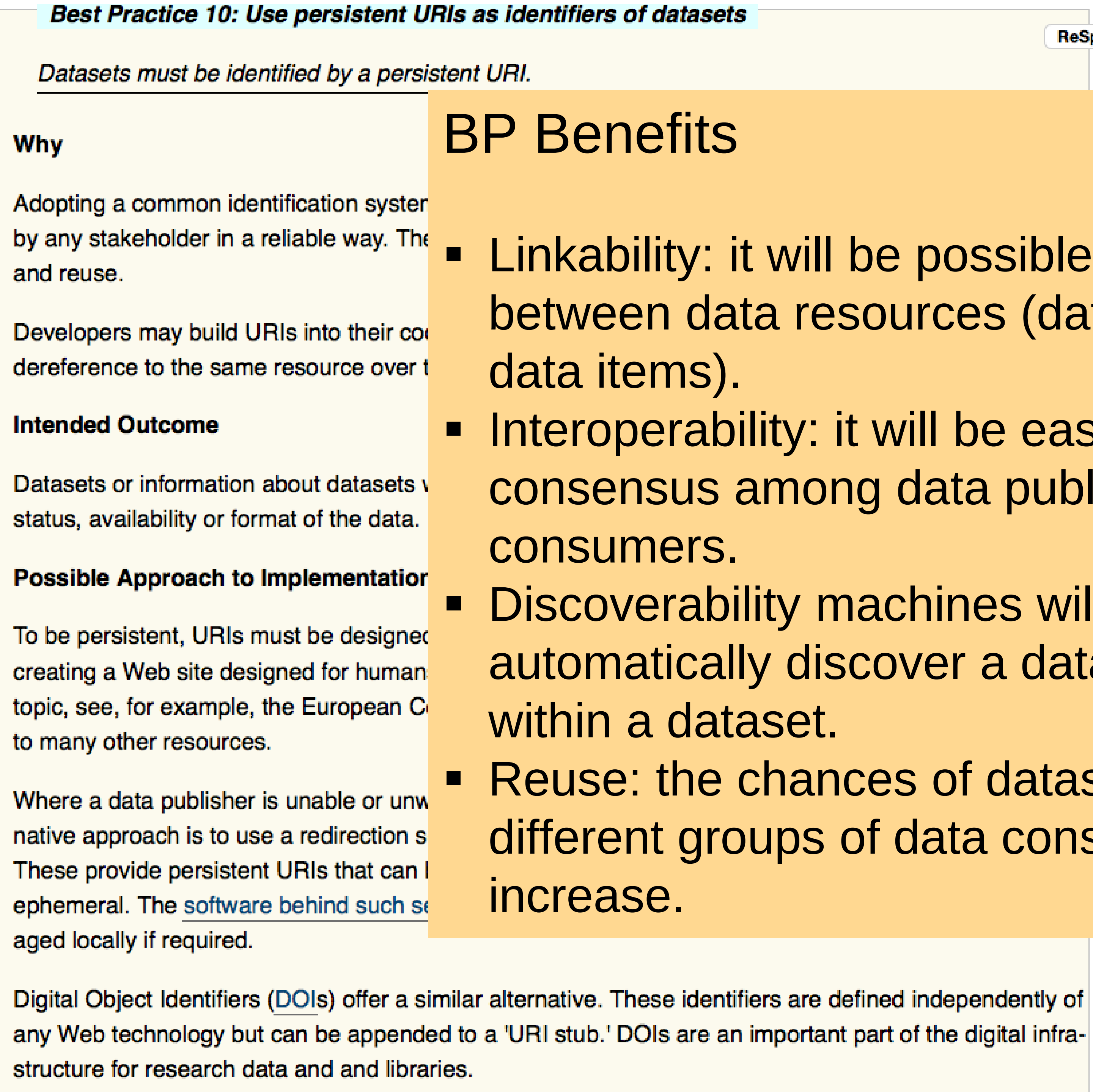
- to provide metadata as part of an H
- to provide metadata as a separate

Possible approaches to provide *machin*

- machine readable metadata may b
it can be embedded in the HTML p
published separately, they should b
nance of multiple formats is best ac
a single source of the metadata.
- when defining machine readable m
ies 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 syster
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 Implementatior

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 s
These provide persistent URIs that can l
ephemeral. The software behind such se
aged 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.
- 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.

Data on the Web Best Practices

1. They are still being developed
2. Publication of the next draft as Candidate Recommendation - April 2016
3. Publication as recommendation - July 2016
4. Feedback is welcome! :)

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

Obrigada!

www.ceweb.br

cburle@nic.br @carolburle

bfl@cin.ufpe.br @bernafarias

April 12th, 2016

nic.br cgi.br

www.nic.br | www.cgi.br