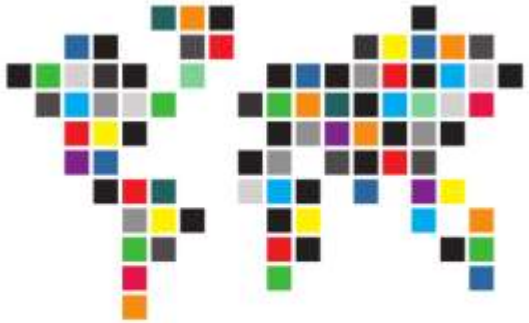


Open Government Partnership

Global Summit • Paris • France 2016

Data on the Web Best Practices: Challenges and Benefits

By Caroline Burle

Topics to be discussed

- Data on the Web Context
  - Data on the Web x Open Data x Linked Data
- 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

Open Data Charter principles

- Open by Default
- Timely and Comprehensive
- Accessible and Usable
- Comparable and Interoperable
- For Improved Governance and Citizen Engagement
- For Inclusive Development and Innovation

Open Data must be on the Web

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.

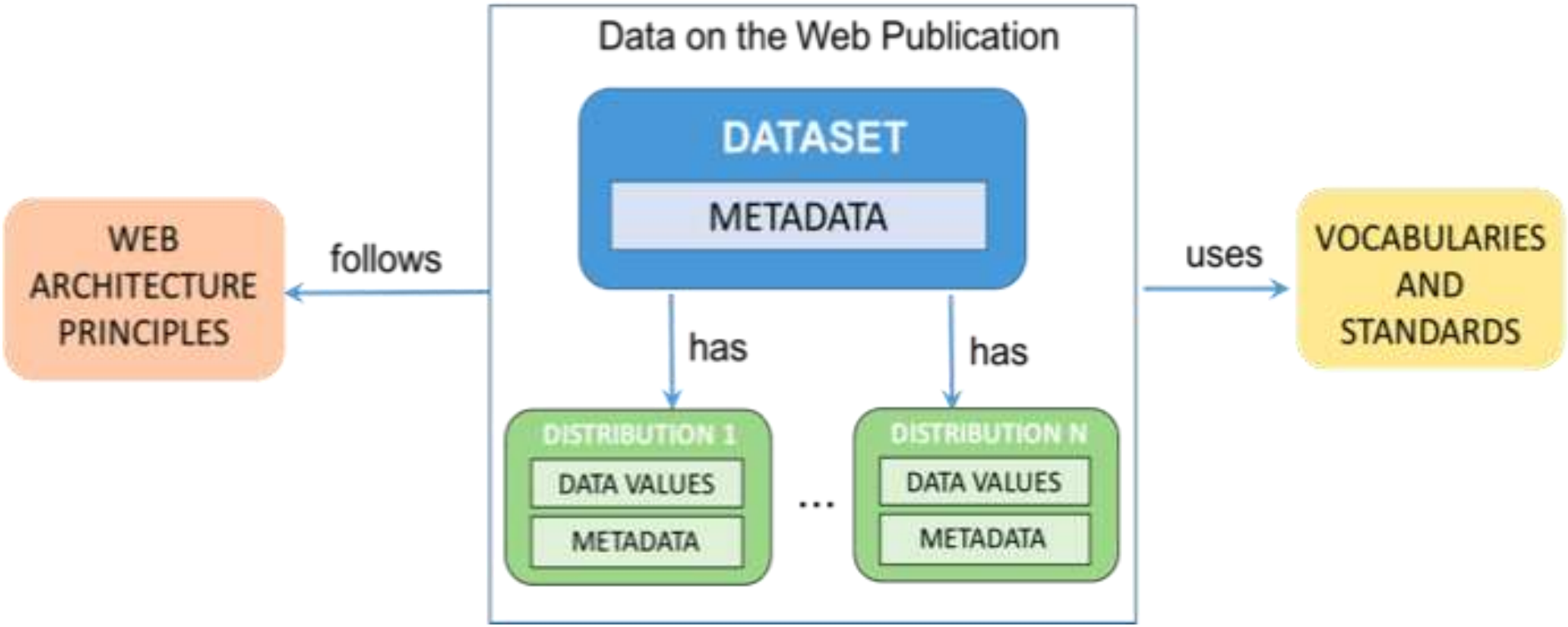
Data on the Web Best Practices Working Group of W3C

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 f**oster trust in the data among developers**, whatever technology they choose to use, increasing the potential for genuine innovation.

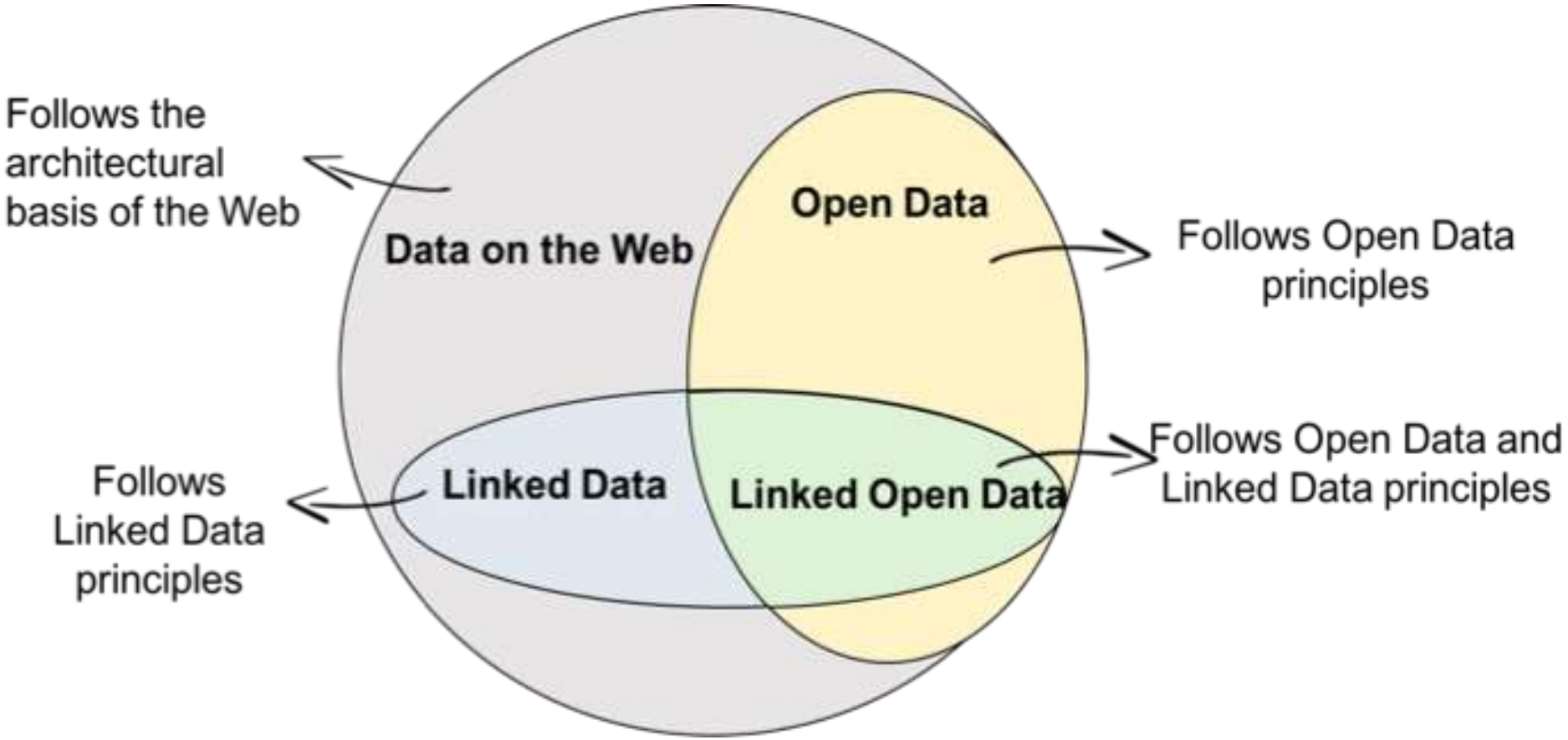
Source: https://www.w3.org/2013/dwbp/wiki/Main_Page

Data on the Web Context



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

Data on the Web x Open Data x Linked Data



Source: http://www.slideshare.net/carolineburle/data-on-the-web-best-practices-challenges-and-benefits

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

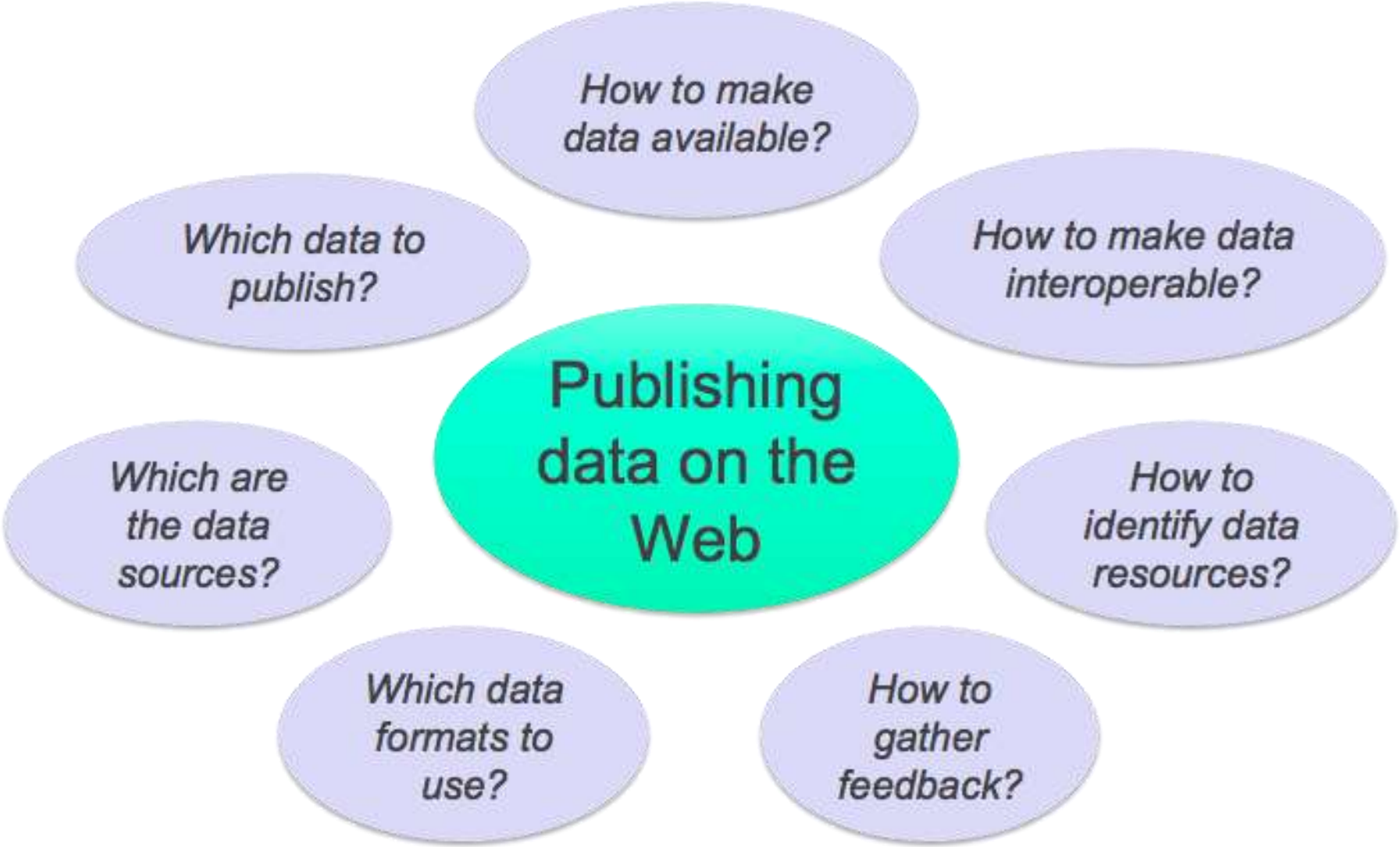
https://www.w3.org/TR/dwbp-ucr/

Data on the Web use cases

Table of Contents

12 challenges and
42 requirements

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 - how to preserve data on the Web?
- Feedback - how to engage users?
- Data Enrichment - adding value to data
- Data Republication - reuse data responsibly

Data on the Web Best Practices

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/

35 Best Practices

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

Best

Best

Best

is nc

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 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 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

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

Discoverability

BP: Provide descriptive metadata
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

Best Practice and Benefits

Best Practice 1: Provide metadata

Metadata must be provided for both human users and computer applications.

Why

Providing metadata is a fundamental requirement when publishing data on the Web because data publishers and data consumers may be unknown to each other. Then, it is essential to provide information that helps human users and computer applications to understand the data as well as other important aspects that describes a dataset or a distribution.

Intended Outcome

Human-readable metadata will enable humans to understand the metadata and machine-readable metadata will enable computer applications, notably user agents, to process the metadata.

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 a serialization format such as Turtle and JSON, or it can be embedded in the HTML page using [HTML-RDFA] or [JSON-LD]. If multiple formats are published separately, they should be served from the same URL using content negotiation. Maintenance of multiple formats is best achieved by generating each available format on the fly based on 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.

Merci!

cburle@nic.br
@carolburle
http://ceweb.br/

Paris, December 9th 2016

This presentation was first presented at the OD Research and can be seen at http://www.slideshare.net/carolineburle/data-on-the-web-best-practices-challenges-and-benefits