Implementation of the Best Practices for Data on the Web in Brazil and Costa Rica

Bernadette Farias Lóscio
bfl@cin.ufpe.br
Center of Informatics - Federal University of Pernambuco
Recife, Brazil

Caroline Burle
cburle@nic.br
Brazilian Network Information Center - NIC.br
São Paulo, Brazil

Newton Calegari
newton@nic.br
Brazilian Network Information Center - NIC.br
São Paulo, Brazil

ABSTRACT

The Best Practices described on the Data on the Web Best Practices (DWBP) document [3] encourages and enables the continued expansion of the Web as a medium for the exchange of data. In this context, this paper focus on two cases of implementing the DWBP. The first one concerns data published by The Regional Center for Studies on the Development of the Information Society (Cetic.br) of The Brazilian Network Information Center (NIC.br). The second use case shows the experience of the Judiciary Department of Costa Rica (Justicia Abierta) with applying the DWBP Recommendation to publish their data on the Web.

KEYWORDS

Data on the Web, Best Practices, Application in Brazil and Costa Rica, Open Data

ACM Reference Format:
Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. 2019. Implementation of the Best Practices for Data on the Web in Brazil and Costa Rica. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316463

1 INTRODUCTION

One of the main goals of the Data on the Web Best Practices (DWBP) [3] is to facilitate interaction between publishers and consumers of data on the Web. A set of 35 Best Practices was created by W3C as a Web Recommendation 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.

In this context, this paper aims to illustrate the use of the DWBP through real use cases. For this, we present the Cetic.br [1] and the Justicia Abierta [2] use cases for applying the set of 35 Best Practices to publish their data on the Web. These two use cases illustrate some of the main challenges faced when applying the DWBP. On the one hand, their evaluation about the DWBP shows that a large number of best practices are easier to implement, on the other hand, both scenarios found difficulties in applying a subset of the best practices.

The remainder of this paper is organized as follows. Section 2 presents an overview of the DWBP. Section 3 presents the Brazil use case. Section 4 presents the Costa Rica use case. Section 5 presents some conclusions.

2 DATA ON THE WEB BEST PRATICES

The Best Practices described on the Data on the Web Best Practices document [3] encourages and enables the continued expansion of the Web as a medium for the exchange of data. In broad terms, data publishers aim to share data either openly or with controlled access. Data consumers (who may also be publishers themselves) want to be able to find, use and link to the data, especially if it is accurate, regularly updated and guaranteed to be available at all times. This creates a fundamental need for a common understanding between data publishers and data consumers. Without this agreement, data publishers' efforts may be incompatible with data consumers' desires.

In this context, it becomes crucial to provide guidance to publishers that will improve consistency in the way data is managed. Such guidance will promote the reuse of data and foster trust in the data among developers, whatever technology they choose to use, increasing the potential for genuine innovation.

The set of Best Practices proposed in [3] was developed to cover the lack of technical guidance for publishing data on the Web. As described in [3], each Best Practice has an intended outcome, which describes "*What it should be possible to do when a data publisher follows the best practice*". In general, the intended outcome is an improvement in the way that a data consumer (human or software) can manipulate a dataset published on the Web. In some cases, the intended outcome reflects an improvement in the dataset itself, which will also result in a gain for the data consumer.

In order to encourage data publishers to adopt the DWBP, the recommendation proposes a number of distinct DWBP benefits, including: comprehension; processability; discoverability; reuse; trust; linkability; access; and interoperability. The benefits are important to engage data publishers in using the Best Practices because they help them to have a better understanding of "what will be possible" when one or more Best Practices are adopted.

Each benefit is associated to one or more specific Best Practices. For example, the benefit "Comprehension" is associated to ten Best Practices, which are related to metadata, data vocabularies, feedback and data enrichment. This means that if a data publisher adopts these Best Practices, the level of comprehensibility will increase, i.e., it will be possible for humans to have a better understanding

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316463

about the data structure, the data meaning, the metadata and the nature of the dataset.

The following list shows the set of best practices linked to the DWBP document:

- 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 a version indicator
- Best Practice 8: Provide version history
- Best Practice 9: Use persistent URIs as identifiers of datasets
- 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
- Best Practice 20: Provide real-time access
- Best Practice 21: Provide data up to date
- Best Practice 22: Provide an explanation for data that is not available
- Best Practice 23: Make data available through an API
- Best Practice 24: Use Web Standards as the foundation of APIs
- Best Practice 25: Provide complete documentation for your 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

3 BRAZIL'S USE CASE

The Center for Studies on the Development of the Information Society (Cetic.br) produces indicators and statistics on the use of information and communication technologies in Brazil since 2005. In 2017, it started the process of providing microdata databases of its sample surveys. Doing this, the Cetic.br allows more people to access and use their data, as well as different devices could recognize and read their data, thus broadening the possibilities of use of the databases published by the Center [1].

Table 1: Assessing the implementation of DWBP on Cetic.br microdata website

Dimension	Best Practice	Applies	Applies immediately
Metadata	BP01	Yes	Yes
	BP02	Yes	Yes
	BP03	Yes	Yes
Data Licenses	BP02	Yes	Yes
Data Provenance	BP05	Yes	Yes
Data Quality	BP06	No	No
Data Versioning	BP07	Yes	Yes
	BP08	Yes	Yes
Data Identifiers	BP09	Yes	Yes
	BP10	No	No
	BP11	No	No
Data Formats	BP12	Yes	Yes
	BP13	No	No
	BP14	Yes	Yes
Data Vocabularies	BP15	Yes	Yes
	BP16	No	No
Data Access	BP17	Yes	Yes
	BP18	No	No
	BP19	Yes	No
	BP20	No	No
	BP21	Yes	Yes
	BP22	Yes	Yes
Data Access APIs	BP23	No	No
	BP24	No	No
	BP25	No	No
	BP26	No	No
Data Preservation	BP27	Yes	Yes
	BP28	No	No
Feedback	BP29	Yes	Yes
	BP30	Yes	Yes
Data Enrichment	BP31	No	No
	BP32	Yes	Yes
Republication	BP33	No	No
	BP34	No	No
	BP35	No	No

In order to improve the quality of their data publication process, the Center for Studies on the Development of the Information Society (Cetic.br) followed the steps described below to apply the DWBP:

(1) Meetings with the teams of Cetic.br, Web Technology Study Center (Ceweb.br), Legal Department and Web Systems Department of NIC.br were held to level the knowledge of these teams on DWBP.
(2) A checklist[1] with the 35 Best Practices was made to understand which of them were applicable right away, which

[1]https://docs.google.com/spreadsheets/d/1yXyd3bnvLWL6KqKZzUbd8_BrB47YG8extgF3eBOHEDw/edit#gid=0

should be postponed and for how long, and the ones that were not applicable.
(3) The DWBP applicable were implemented on Cetic.br Microdata Website[2]

An important step in the process of applying the DWBP concerns the meetings with the teams of Cetic.br, Ceweb.br, Legal Department and Web Systems Department of of NIC.br to level the knowledge of these teams on DWBP. After the meetings, the necessary adaptations - both in the implementation of the BPs and in the processes of the research of Cetic.br -, were implemented. This was also a first step to apply the DWBP and to become part of Cetic.br's research process and data publication.

Table 1 shows the result of the checklist, meaning which BPs were applicable, which ones were not applicable and which ones should be adapted before being implemented.

The checklist presented in the Table 1 shows that 18 Best Practices could be applied immediately and the BP19 "Use content negotiation for serving data available in multiple formats" should be applied afterwards because the Web System Department of NIC.br needed more time to implement such BP.

Figure 1: Cetic.br microdata website

Most of the 16 BPs not applicable would not be implemented because of specificities of the research methodology and data production. For instance, BP20 "Provide real-time access" could not be implemented because of the data collection process, which does not allow to publish data in real time. Another example of BPs that could not be implemented are the ones in the Dimension of Republication: BP33 "Provide Feedback to the Original Publisher"; BP34 "Provide Feedback to the Original Publisher"; and BP35 "Cite the Original Publication". Regarding that Cetic.br is the data producer and does not republishes data from other statistical institutes.

According to [1], one of the lessons learned from the team involved in this process was that, despite being technical recommendations, it is not the technical character that limits its implementation. The main barrier is the engagement around publication and the perception of importance and benefits - for publishers and for data consumers - to implement the BPs and to publish and share the Center's databases. In this sense, the importance of involving people from different areas is highlighted, so as to promote an engagement and alignment between all the parties involved. The Figure 1 shows part of the Cetic.br microdata website front page.

4 COSTA RICA'S USE CASE

In 2017, the Judiciary Department of Costa Rica was supported by the The Trust for the Americas Foundation to create the Justicia Abierta (Open Justice) and open its data [2].

Similar to the Cetic.br, the Judiciary Department of Costa Rica followed the steps described below to apply the DWBP:

(1) The Web Technology Study Center (Ceweb.br) team ministered a course on publishing data on the web - based on the Data on the Web Best Practices (DWBP) - for twenty employees of the Government of Costa Rica, in February of 2017.
(2) The Open Justice Department created an interdisciplinary group to add knowledge on publishing data on the Web.
(3) The Justice Department team conducted a public consultation in partnership with the Abriendo Datos initiative to obtain from the Costa Rican society which data they should prioritize. They then collected, cleaned and checked those data.
(4) A checklist with the 35 Best Practices was made to understand which of them were applicable right away, which should be postponed and for how long, and the ones that were not applicable.
(5) The DWBP applicable were implemented on Justicia Abierta Website[3].

Table 2 shows the result of the checklist, meaning which BPs were applicable, which ones were not applicable and which ones would need to be adapted before implemented. The checklist shows that 16 Best Practices were to be applied immediately and 06 BPs should be applied afterwards. Technical limitations were the main reason for postponing the adoption of these 06 BPs.

The biggest challenge reported by the Justicia Abierta leader was the lack of knowledge on the part of public officials involved in the choice of those data and also some resistance by some guardians of the data. According to him, these difficulties were overcome by means of conversations and training that took place with the interdisciplinary group[4].

At the end, after the internal process have been established, of the 35 DWBP, the Justice Department of Costa Rica applied 17 BPs. It is noteworthy that, prior to step 1, only 3 BPs were adopted by the Costa Rica Justice Department.

[2]https://www.cetic.br/microdados/

[3]http://datosabiertospj.eastus.cloudapp.azure.com/

[4]http://ceweb.br/media/docs/publicacoes/10/DWBP_Condatos_Caroline%20Burle_Carlos%20Morales.pdf

Table 2: Assessing the implementation of DWBP on the Justice Department website

Dimension	Best Practice	Applies	Applies immediately
Metadata	BP01	Yes	Yes
	BP02	Yes	Yes
	BP03	Yes	Yes
Data Licenses	BP04	Yes	Yes
Data Provenance	BP05	Yes	Yes
Data Quality	BP06	Yes	Yes
Data Versioning	BP07	Yes	Yes
	BP08	Yes	Yes
Data Identifiers	BP09	Yes	No
	BP10	No	No
	BP11	Yes	No
Data Formats	BP12	Yes	Yes
	BP13	Yes	No
	BP14	Yes	Yes
Data Vocabularies	BP15	Yes	Yes
	BP16	Yes	No
Data Access	BP17	Yes	Yes
	BP18	Yes	No
	BP19	No	No
	BP20	No	No
	BP21	Yes	Yes
	BP22	No	No
Data Access APIs	BP23	No	No
	BP24	No	No
	BP25	No	No
	BP26	No	No
Data Preservation	BP27	Yes	No
	BP28	Yes	Yes
Feedback	BP29	Yes	Yes
	BP30	No	No
Data Enrichment	BP31	No	No
	BP32	Yes	Yes
Republication	BP33	No	No
	BP34	No	No
	BP35	No	No

The Figure 2 shows *Justicia Abierta* website open data front page.

5 DISCUSSION

This paper presented two real use cases of the DWBP application: the Cetic.br [1] and the Justicia Abierta [2] use cases. Both use cases adopted a similar methodology to implement the DWBP, including team meetings as well as a detailed evaluation of the DWBP.

Both Cetic.br and *Justicia Abierta* use cases demonstrated the importance of gathering all people involved in the process of publishing data. In the case of Cetic.br, even though the data was produced by the Center, holding meetings with the other departments involved such as the Legal Department and Web System Department of NIC.br, as well as with Ceweb.br to help the explanation about the DWBP, was very important to the process. In the case of

Figure 2: Costa Rica *Justicia Abierta* website

Open Justice, conversations and trainings that took place with the interdisciplinary group were significant to overcome the difficulties on lack of knowledge and resistance by some guardians of the data.

It is important to note that in both cases some Best Practices could be implemented immediately: 1 to 5 - about Metadata, Data Licenses and Data Provenance; 7 and 8 - about Data Versioning; 12 and 14 - about Data Formats; 15 - about Data Vocabularies; 17 and 21 - on Data Access; 29 - about Feedback; and 32 - about Data Enrichment. This shows that some of the proposed best practices can be easily adopted without major modifications on the data publication process. On the other hand, neither were able to implement the BPs 10 - about Data Identifiers; 20 - on Data Access; 23 to 26 - about Data Access APIs; 31 - about Data Enrichment; and 33 to 35 - about Republication. For implementing these BPs, more technical or advanced knowledge becomes necessary.

Although neither use cases implemented all the Data on the Web Best Practices, both use cases showed that it is possible to implement some of them without great effort. Furthermore, according to both teams, understand the DWBP as a guide and not as rules to be fulfilled is of great importance. We understand that both Brazil's and Costa Rica's use cases may be a reference to other data publishers of how to use the DWBP to facilitate interaction between data publishers and consumers.

REFERENCES

[1] Tatiana Jereissati Javiera Medina Macaya Beatriz Rossi Corrales, Caroline Burle. 2018. A importância das Boas Práticas para Dados na Web e o caso do Cetic.br. *GT.ADM09 Medición Y Evaluación En Apertura De Datos: Desafíos, Retos, Experiencias.* (2018).

[2] Caroline Burle. 2018. Implementação das boas práticas de dados na web: o caso da Costa Rica. (2018). [Online; Accessed 21 Jan 2019].

[3] Bernadette Farias Lóscio, Caroline Burle, and Newton Calegari. 2017. Data on the Web Best Practices. https://www.w3.org/TR/dwbp/. *W3C Recommendation* (2017). [Online; accessed 13 Jan 2019].