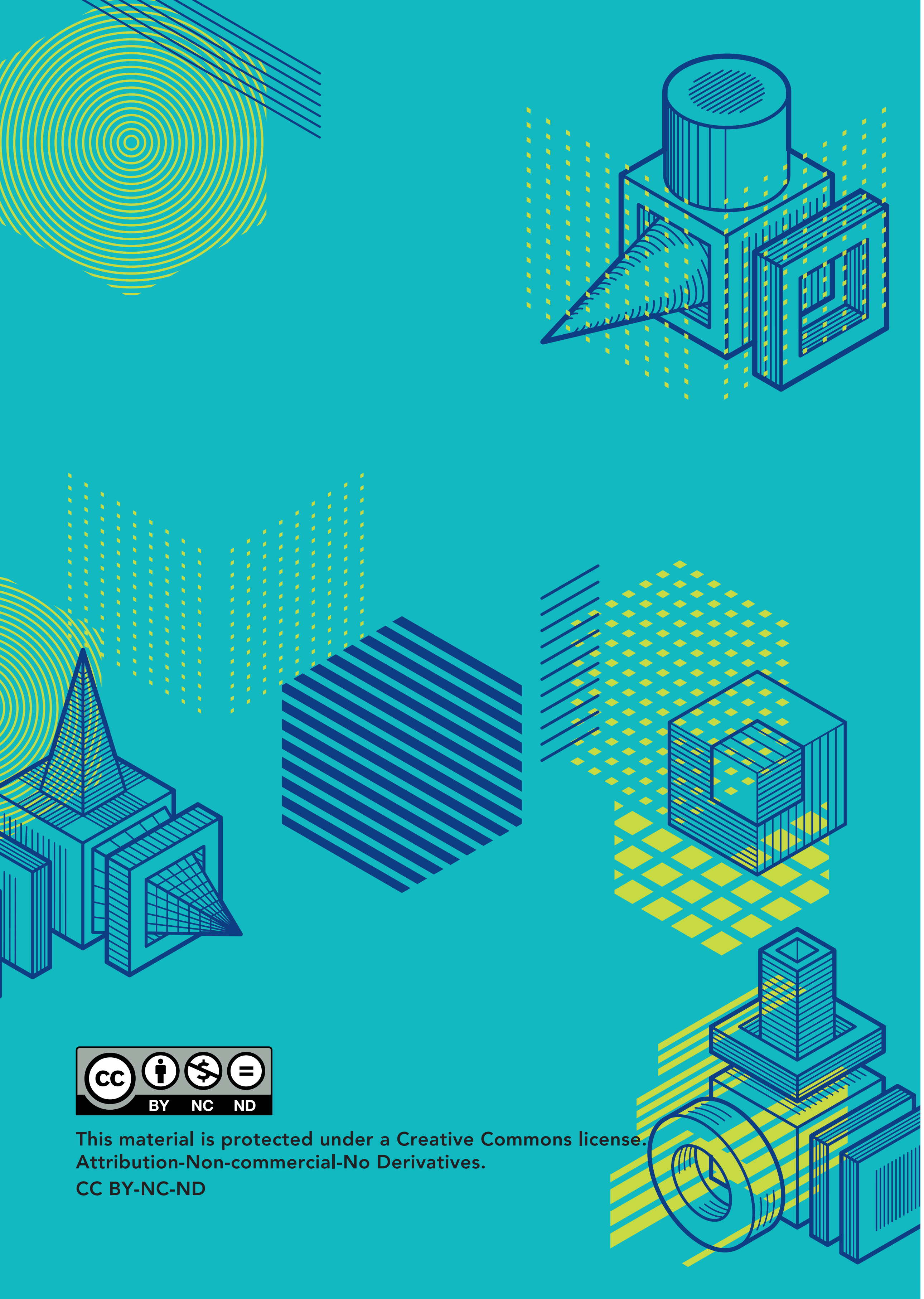
FUNDAMENTALS FOR PUBLISHING DATA ON THE WEB

ceweb.br nic.br cgi.br

TRUST
THE TRUST FOR THE AMERICAS

OAS More rights for more people



FUNDAMENTALS FOR PUBLISHING DATA ON THE WEB

Bernadette Farias Lóscio (UFPE)
Caroline Burle (Ceweb.br/NIC.br)
Marcelo Iury S. Oliveira (UFRPE)
Newton Calegari (Ceweb/NIC.br)

CGI.br
Brazilian Internet
Steering Committee
2018

This material was developed by **the Web Technologies Study Center of the Brazilian Network Information Center (Centro de Estudos sobre Tecnologias Web do Núcleo de Informação e Coordenação do Ponto Br) (Ceweb.br/NIC.br)** within the scope of the Project "From an Open Government to an Open State", carried out by **The Trust for the Americas** of the **Organization of American States (OAS)** through the financing of the US Embassy in San Jose, Costa Rica.

ceweb.br nic.br cgi.br

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR – NIC.BR

Chief Executive Officer: Demi Getschko
Administrative and Finance Director: Ricardo Narchi
IT and Services Director: Frederico Neves
Director for Special Projects and Development: Milton Kaoru Kashiwakura
Consulting Director for CGI.br Activities: Hartmut Richard Glaser

WEB TECHNOLOGY STUDY CENTER – CEWEB.BR

Organization: Beatriz Rossi Corrales

Technical Team: Amanda Marques, Beatriz Rossi Corrales, Caroline Burle, Diogo Cortiz, Mariana Frizanco, Newton Calegari, Reinaldo Ferraz and Selma de Morais

Revision: Caroline Burle, Bernadette Farias Lóscio and Beatriz Rossi Corrales

Production: Caroline D'Avo (Communication NIC.br) and Everton Rodriguez (Communication NIC.br)

Graphic and Illustration Project: Giuliano Galvez **(Communication NIC.br)**

AUTHORS

Bernadette Farias Lóscio
Information Technology Center – Universidad Federal de Pernambuco (UFPE)
blf@cin.ufpe.br

Caroline Burle
Web Technologies Study Center (Ceweb.br)
Brazilian Network Information Center (NIC.br)
cburle@nic.br

Marcelo Iury S. Oliveira
Serra Talhada Academic Unit – Universidad Federal de Pernambuco (UFPE)
marcelo.iury@ufrpe.br

Newton Calegari
Web Technologies Study Center (Ceweb.br)
Brazilian Network Information Center (NIC.br)
newton@nic.br

TABLE OF CONTENTS

INTRODUCTION

Since its beginnings, the Web has stood out as an important means for the exchange of information. In this scenario with a great amount of data available on the Web, there are two roles which deserve being pointed out: publishers and consumers of data. In general terms, the goal of data publishers is the publishing and exchange of data, with free or controlled access, while data consumers (who may also be publishers) want to use such data for generating useful and pertinent information as well as generating new data.

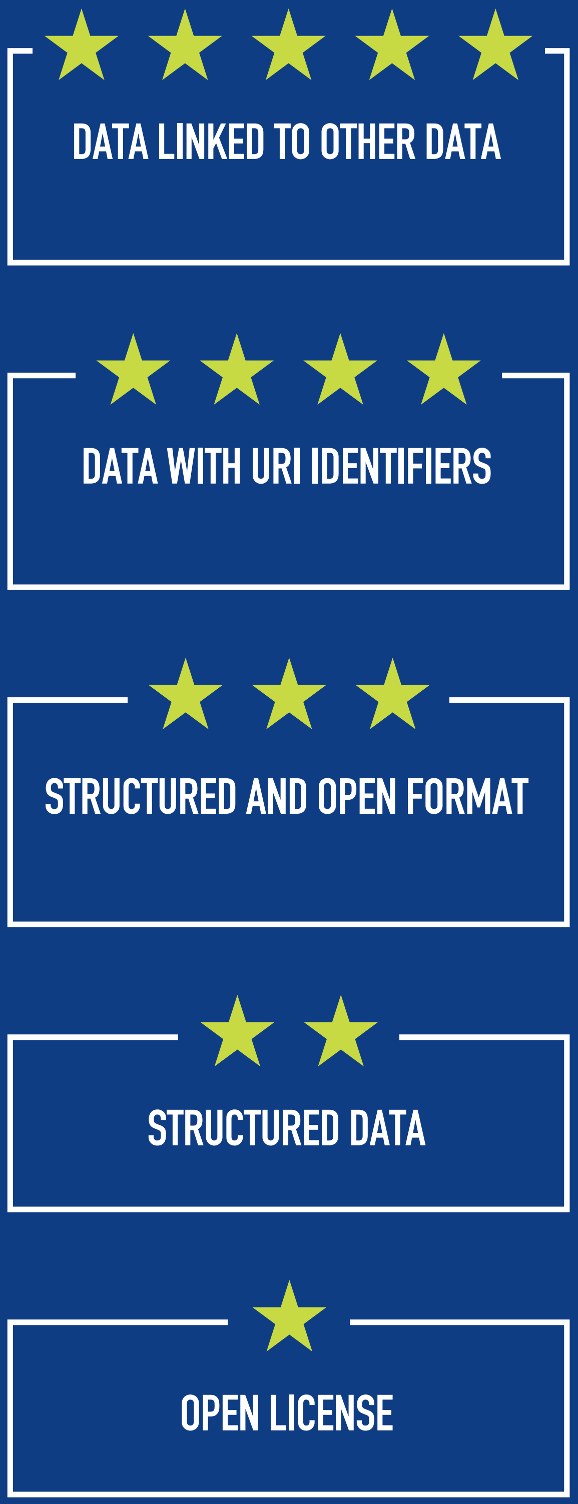
It is important to point out that the interest for publishing data on the Web is not something new (BERNERS-LEE; CONNOLLY; SWICK, 1999 and ABITEBOUL; BUNEMAN; SUCIU, 2000). However, during the last years, this interest has been characterized by the publishing of data in order to promote its exchange and reutilization. Therefore, providing access to data is simply not enough. In general, data must be published in order that consumers may easily comprehend and use them and in order that, moreover, these are available in formats which may be easily processed by applications. However, the heterogeneity of data and lack of standards for the description and access to datasets turn the process of publishing, sharing and consuming data into a complex task. Within this context, this article discusses the fundamentals related to the publishing of data on the Web, making reference to relevant aspects, including, among them, the concepts of Open Data, Linked Data, Life Cycle of Data on the Web and Data on the Web Best Practices.

OPEN DATA

According to the Open Knowledge Foundation (2012), Open Data is any data that may be freely used, reused and redistributed by any person. As a result, open data consists in publishing and diffusion of information through the Internet, shared in open formats, legible by machines and which may be freely reused by the society in an automatized manner. In other words, the opening of data seeks avoiding a control and restriction mechanism over published data, allowing that both individuals and legal entities may freely exploit such data (ISOTANI; BITTENCOURT, 2015). Data is considered open when it has the following characteristics (OPEN KNOWLEDGE, 2012):

I. Availability and access: data must be available in its entirety. It must be in a convenient and modifiable format;
II. Reuse and redistribution: data must be supplied in conditions of reuse and redistribution and it must be combinable with other data;
III. Universal participation: everyone may use, reuse and redistribute the data with no restrictions regarding areas, persons or groups.

Open data may be classified according to a scale based on stars proposed by Tim Berners-Lee (BERNERS-LEE, 2006). According to this classification, shown in Figure 1, data published on the Web in any format (image, table or document) and associated to a license which allows its use and reuse under no restrictions is classified with a score of 1 Star. Despite of being a progress already, 1-star data must be used manually or through extractors specifically built for accessing data.



★★★★★

DATA LINKED TO OTHER DATA

★★★★

DATA WITH URI IDENTIFIERS

★★★

STRUCTURED AND OPEN FORMAT

★★

STRUCTURED DATA

★

OPEN LICENSE

Figure 1: This illustration is based on the diagram proposed by Tim Berners-lee (2006).

From the moment in which data is published in a format that may be automatically processed by any software (e.g. an Excel spreadsheet rather than an image), data is classified with a 2-star score. On one hand, this may simplify the work of the data consumer; however, on the other hand, it may slightly complicate the publishing task.

Data is classified with a 3-star score when published in non-proprietary file formats (e.g. CSV rather than Excel). Again, data publishing in open formats may imply additional costs for publishers. This happens when the source format is different from the format implemented for the publishing and it is necessary to convert the data and keep coherence between the original data source and the data published in open format.

When data receives a unique identification and is linked to other data, these data may be classified with a 4-star score. The creation of links between data allows them being part of a broader network of open and linked data (BIZER; HEATH; BERNERS-LEE, 2009). Finally, data is classified with a 5-star score if linked to other data already available on the Web. In this case, it is necessary to identify data that represents the same concept in order to establish links between them.

Based on the open data movement, governments from different countries are using the Web as a means for publishing data and information about their administrations. Such Open Government Open Data may be easily found in the so-called Open Data Portals, which offer a friendlier interface for cataloguing and accessing data. Some of the most relevant examples of already-consolidated open data portals include the US[1] portal and the UK[2] portal. Several European countries, such as France[3] and Netherlands[4], as well as some Latin American countries, such as Chile[5] and Uruguay[6], also have open government data portals. In the case of Brazil, the open data portal[7] was launched at the beginning of 2012 through an initiative led by the Ministry of Planning.

[1] http://data.gov

[2] http://data.gov.uk

[3] http://data.gouv.fr

[4] http://dataoverheid.nl

[5] http://datos.gob.cl

[6] http://datos.gub.uy

[7] http://dados.gov.br

The open data initiative implemented by governments has been encouraged by the seek of transparency, cooperation and participation of the society and community (GOLDSTEIN; DYSON, 2013). For purposes of reaching consensus on the necessary requirements for classifying an open database, the Open Government Working Group developed eight principles of open government data (TAUBERER; LESSIG, 2007):

Complete: all public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.

Primary: data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.

Timely: data is made available as quickly as necessary to preserve the value of the data.

Accessible: data is available to the widest range of users for the widest range of purposes.

Machine-readable: data is reasonably structured to allow automated processing.

Non-discriminatory: data is available to anyone, with no requirement of registration.

Non-proprietary: data is available in a format over which no entity has exclusive control.

License-free: data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Open government data refers to diverse matters and may involve both data on the expenses and revenues of the government and data on school census, tourist spots, consumer complaints, service demands and many others. In general, the available data comes from routine activities carried out within governmental entities, such as ministries and secretary’s offices.

Once the governmental data is available in open format, it is expected that it be used in the development of accessible applications that may be easily used by both citizens and the government itself. Applications offer means for analyzing data through filters and also allow visualizing data in a simple and creative manner. There are already different applications and visualization tools available on the Web, mainly as a result of competitions and hackathons developed for the diffusion and promotion of open data portals.

LINKED DATA

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web in order to create a "Web of Data" (BIZER; HEATH; BERNERS-LEE, 2009). The Web of Data gives rise to countless opportunities for the semantic integration of data itself, encouraging therefore the development of new types of applications and tools, such as browsers and search engines (ISOTANI; BITTENCOURT, 2015).

For purposes of having a better understanding on the Web of Data, a parallel relationship may be established with the Web of Documents (i.e. the current Web) and the Web of Data. The former uses the standard HTML for accessing data while, in the latter, data is accessed from a standard RDF (ISOTANI; BITTENCOURT, 2015). The Web of Documents is based on a set of standards that includes: a global and unique identification mechanism, URIs (Uniform Resource Identifiers), a universal access mechanism, the HTTP, and a standard content representation format, the HTML. Similarly, the Web of Data is also based on a series of standards, including, among these: the same universal identification and access mechanism used by the Web of Documents (URIs and HTTP, respectively), a standard model for the representation of data, the RDF, and a query language for accessing data, the SPARQL language (ISOTANI; BITTENCOURT, 2015).

The principles of Linked Data were initially introduced by Tim Berners-Lee (2006) and are summarized in four basic principles:

I. Use URIs as names for things.
II. Use HTTP URIs so that people can look up those names.
III. When someone looks up a URI, provide useful information.
IV. Include links to other URIs. so that they can discover more things.

The first principle refers to the use of URIs for identifying not only Web documents and digital contents but also things from the real world and abstract concepts that must be represented in RDF format.

The second principle refers to the use of HTTP URIs for identifying the things and abstract concepts defined by Principle 1, allowing that such URIs may be dereferenced over an HTTP protocol. In this context, dereferencing is the process of retrieving a representation of a resource identified by an URI, where the resource may have several representations as HTML, RDF, XML documents or others.

In order that a wide range of applications may process the data available on the Web, there must be an agreement regarding a standard format to be used for making data available. The third principle of Linked Data refers to the use of RDF as a model for publishing structured data on the Web (CYGANIAK; WOOD; LANTHALER, 2014). The RDF allows describing resources, setting up software agents for exploring data automatically, often adding, interpreting or combining data.

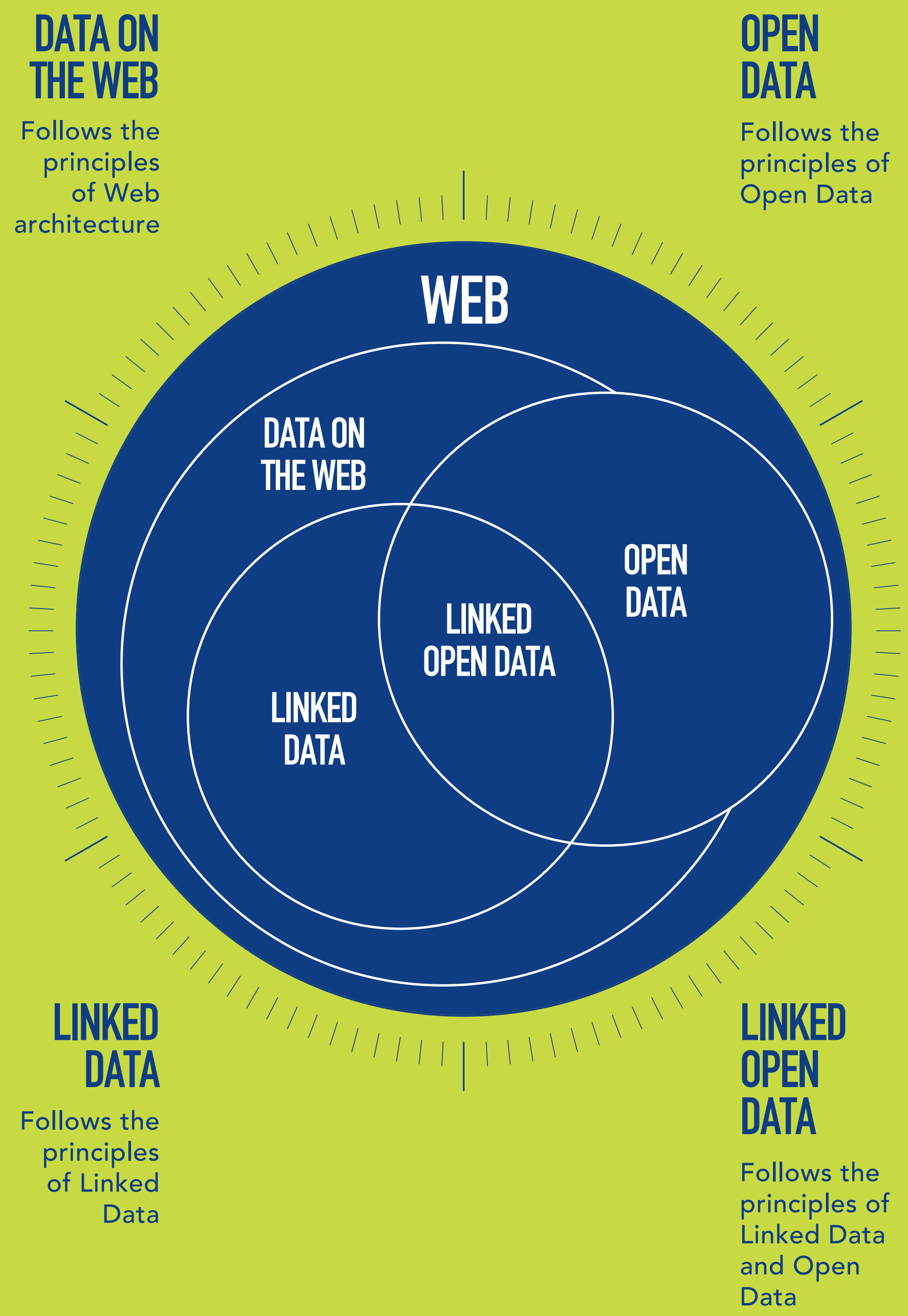
The fourth principle refers to the use of links for connecting not only Web documents but any other type of resource. For example, a link may be created between a person and a place, or between a location and a company. In comparison to the classic Web in which hyperlinks are mostly "untyped", hyperlinks that connect resources in a Linked Data context are capable of describing the relationship between them. In the Linked Data context, hyperlinks are called RDF links in order to differentiate them from hyperlinks existing on the conventional Web (HEATH; BIZER, 2011).

It is important to point out that, nowadays, there is a large volume of linked open data available on the Web. For example, emphasis may be done in the set of open data published by the LOD[8] project. As previously mentioned, Linked Data contributes to generating a Web of Data, therefore, it is the preferred option for publishing data on the Web. In this context, the W3C Government Linked Data Working Group proposed a set of Best Practices for publishing Linked Data in order to provide guidelines which facilitate the access and reuse of open government data[9].

http://lod-cloud.net [8]
http://www.w3.org/TR/ld-bp [9]

DATA ON THE WEB

Data on the Web is the general term that may be used upon making reference to the data published according to the architectural basis of the Web (JACOBS; WALSH, 2004). As shown in Figure 2, data on the Web may be classified as Open Data (PIRES, 2015), Linked Data and Linked Open Data (BERNERS-LEE, 2006). According to the Open Data Charter, "open data is digital data that is made available with the technical and legal characteristics necessary for it to be freely used, reused and redistributed by anyone, anytime, anywhere". Since the Web is the most adequate means making open data available, such data is also data on the Web. A part of the data currently available on the Web follows these principles and is classified as Linked Data. Finally, when a set of data is published on the Web following both the principles of Open Data and the principles of Linked Data, such data may be classified as Linked Open Data.


DATA ON THE WEB
Follows the principles of Web architecture
OPEN DATA
Follows the principles of Open Data
LINKED DATA
Follows the principles of Linked Data
LINKED OPEN DATA
Follows the principles of Linked Data and Open Data
WEB
DATA ON THE WEB
OPEN DATA
LINKED OPEN DATA
LINKED DATA

It is important to make emphasis on the fact that not all datasets published on the Web are openly shared, meaning that a great part of the data published on the Web is "closed". Upon defining the data publishing policy and the circumstances under which data must be published, data publishers must take into account the security, commercial sensitivity and, most of all, the privacy of persons.

Figure 2:
Intersection of Data on the Web,
Open Data and Linked Data.
Source: the authors

LIFE CYCLE OF DATA ON THE WEB

The process of publishing and consuming data on the Web implies different phases going from the selection and publication of data to the use of data and the feedback on the used data. This set of phases comprised in the data publishing and consumption process is called Life Cycle of Data on the Web. Figure 3 shows the phases of Life Cycle of Data on the Web that are briefly described below.

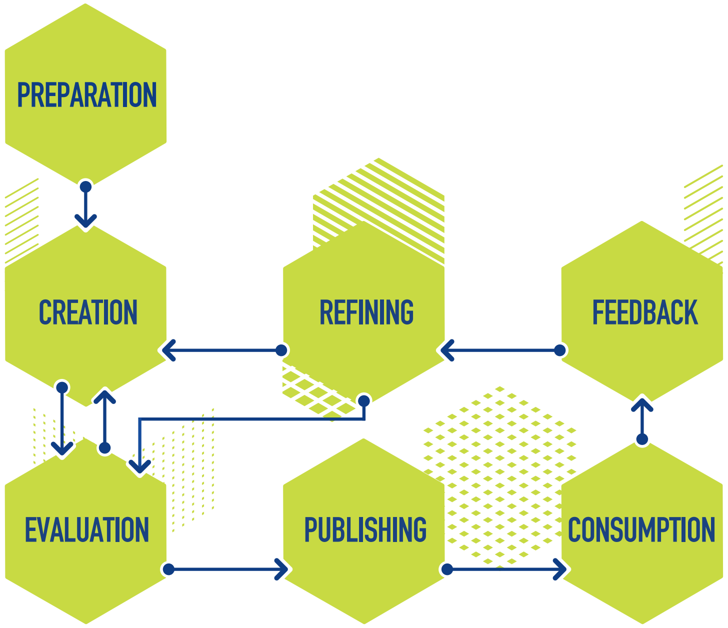


Figure 3:
Life Cycle of Data on the Web
Source: the authors

Preparation: this phase comprises from the moment in which the intention of publishing data arises to the selection of the data to be published. It is important to bear in mind that there are no rules for establishing the priority of the data to be published, but it is always important to take into consideration the relevance of data. In other words, those data having great use potential should have priority upon taking the decision. Therefore, whenever possible, it is important to previously consult with potential data consumers in order to identify the relevance of data.

Preparation: t

Creation: refers to the moment in which data is created, i.e. includes the phase comprised from the extraction of data from already existing data sources to its transformation to the appropriate format in order to be published on the Web. The creation phase also includes the selection of the data formats to be used for publishing data and metadata. Moreover, it is also appropriate taking into consideration the publishing of data in different formats in order to minimize the consumers' needs for transforming such data.

Evaluation: this phase refers to the evaluation of data prior to its publishing. It is important that experts may evaluate the data in order to detect inconsistences or errors therein, as well as for pointing out which confidential data should not be published, for example. Data should only be available for publishing after being carefully evaluated. Whenever necessary, data may return to the prior phase in order to solve any issue detected by experts.

Publishing: refers to the moment in which data is made available to the public on the Web. Data catalog tools may be used for these purposes, such as CKAN[10] and Socrata[11]. Application Program Interfaces (APIs) may also be used, providing easy access to published data or webpages, for example. In any case, data publishers must offer all necessary information in order that consumers may easily access the data. Moreover, it is important to guarantee that data will be updated on a predetermined basis, updates which must be available together with the data.

Consumption: refers to the moment in which data is used for creating visualizations, such as graphics and heatmaps, as well as for applications that allow cross-examining and analyzing data. This Life Cycle phase is directly related to the data consumer, which may be either a large company interested in using the data available on the Web for improving its products and services or a developer interested in using the data for creating an application that improves the quality of life in his city.

Feedback: this phase includes the moment in which consumers must share their comments on previously used data and metadata. This phase has fundamental relevance since the feedback shared by consumers will allow identifying improvements and making corrections to previously published data. Moreover, this communication channel between consumers and data publishers also facilitates the identification of new relevant data that must be given priority at the time of selecting the new data to be published.

Refining: this phase includes all activities related to additions or updates of data that has been already published. It is very important to guarantee the maintenance of previously published data in order to offer greater security to consumers. Maintenance may be performed according to the feedback shared by consumers or, otherwise, new versions may be developed in order to guarantee that data does not become obsolete. Therefore, it is important to correctly manage the different versions of data and guarantee that consumers have access to the correct version.

[10] http://ckan.org

[11] http://www.socrata.com

Regarding the parties involved in the Life Cycle of Data on the Web, these may have two major roles: data publishers and data consumers. The role of data publishers may be performed by several parties, who are responsible for carrying out activities such as the creation of metadata, creation and publishing of data. Data consumers are those who receive and consume data. It is important to point out that data consumers may also be data publishers, since consumers may improve and refine data in order to offer such data again to the community. It is also important to point out that the Life Cycle mentioned herein does not require following all steps before starting a new iteration.

DATA ON THE WEB BEST PRACTICES

The Data on the Web Best Practices (DWBP) described in the Recommendation of the W3C of Lóscio, Burle and Calegari (2017) were developed for encouraging and enabling the continued expansion of the Web as a means for the exchange of data. In general terms, data publishers seek sharing data openly or under controlled access. Data consumers seek being able to fund, use and establish links to the data, especially if it is accurate and updated data which permanent availability may be guaranteed. Therefore, it is essential that there is a common understanding between data publishers and data consumers. Otherwise, data publishers' efforts could be incompatible with consumers' needs.

In this context, it is crucial to provide guidance to publishers in order to improve the consistency in the manner in which data is managed. It is expected that such guidance encourages the reuse of data and increases the confidence on data by developers, despite of the technology they use, in order to therefore increase the potential for genuine innovation. The set of Best Practices proposed by Lóscio, Burle and Calegari (2017) was developed for purposes of providing technical guidance for publishing data on the Web, contributing to the improvement of the relationship between data publishers and data consumers.

These Best Practices comprise different challenges and requirements related to the publishing and consumption of data, including, among them, data formats, access to data identifiers, data vocabularies and metadata. On one hand, each best practice is focused on at least one of the requirements identified in the document of use cases related to Data on the Web Best Practices (LEE; LÓSCIO; ARCHER, 2015), therefore, the relevance of the best practice is evident based on such requirements. On the other hand, each requirement is addressed by at least one best practice.

As described in Lóscio, Burle and Calegari (2017) and shown in Table 1, each best practice has an intended outcome which describes "what 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) may manipulate a dataset published on the Web. In some cases, the intended outcome reflects an improvement in the dataset itself, which will also benefit the data consumer".

The Best Practices proposed for publishing and using data on the Web refer to datasets, i.e. "a collection of published data, administered by a single agent and available for access or download in one or more formats" (MAALI; ERICKSON, 2014). Through the term data, we refer to "known facts that may be recorded and have an implicit meaning" (ELMASRI; NAVATHE, 2010). As described in Figure 4, data is published in different distributions which are a specific physical form of a dataset. These distributions enable the exchange of data on a large scale, allowing that datasets may be used by different groups of data consumers. In other words, "a person or group accesses, uses and potentially performs the post-processing of data" (STRONG; LEE; WANG, 1997), without taking into consideration its purpose, recipient, interest or license. Taking into account this heterogeneity and the fact that data publishers and consumers may be unknown to each other, it is necessary to provide certain information on datasets and distributions which may also contribute to their reliability and reuse, such as: structural metadata, descriptive metadata, access to information, information on data quality, data provenance information, licensing information and use information.

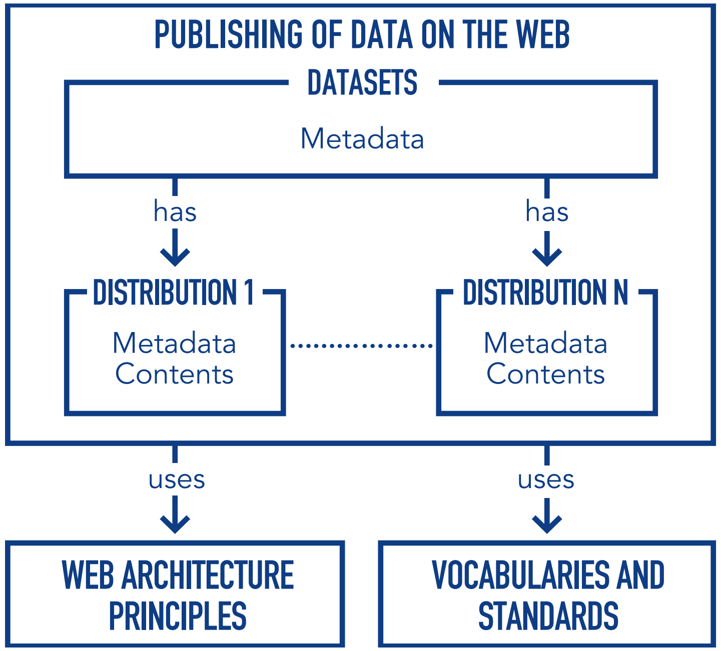


Figure 4: Context of publishing of data on the Web. Source: Lóscio, Burle and Calegari (2017)

Finally, a relevant matter about the publishing and exchange of data on the Web relates to the architectural basis of the Web (JACOBS; WALSH, 2004). In this sense, a relevant aspect is the principle of identification, which indicates that URIs must be used for identifying resources. In our context, a resource may be a complete dataset or a specific element from a specific dataset. All resources must be published with stable URIs in order that these may be referenced and connections may be done between two or more resources through the URIs.

DATA ON THE WEB BEST PRACTICES WITH THEIR RESPECTIVE INTENDED OUTCOMES

BP1 PROVIDE METADATA

Humans will be able to understand the metadata and computer applications, notably user agents, will be able to process it.

BP2 PROVIDE DESCRIPTIVE METADATA

Humans will be able to interpret the nature of the dataset and its distributions, and software agents will be able to automatically discover datasets and distributions.

BP3 PROVIDE STRUCTURAL METADATA

Humans will be able to interpret the schema of a dataset and software agents will be able to automatically process distributions.

BP4 PROVIDE DATA LICENSE INFORMATION

Humans will be able to understand data license information describing possible restrictions placed on the use of a given distribution, and software agents will be able to automatically detect the data license of a distribution.

BP5 PROVIDE DATA PROVENANCE INFORMATION

Humans will know the origin and history of the dataset and software agents will be able to automatically process provenance information.

BP6 PROVIDE DATA QUALITY INFORMATION

Humans and software agents will be able to assess the quality and therefore suitability of a dataset for their application.

BP7

PROVIDE A VERSION INDICATOR

Humans and software agents will easily be able to determine which version of a dataset they are working with.

BP8

PROVIDE VERSION HISTORY

Humans and software agents will be able to understand how the dataset typically changes from version to version and how any two specific versions differ.

BP9

USE PERSISTENT URIS AS IDENTIFIERS OF DATASETS

Datasets or information about datasets will be discoverable and citable through time, regardless of the status, availability or format of the data.

BP10

USE PERSISTENT URIS AS IDENTIFIERS WITHIN DATASETS

Data items will be related across the Web creating a global information space accessible to humans and machines alike.

BP11

ASSIGN URIS TO DATASET VERSIONS AND SERIES

Humans and software agents will be able to refer to specific versions of a dataset and to concepts such as a 'dataset series' and 'the latest version'.

BP12

USE MACHINE-READABLE STANDARDIZED DATA FORMATS

Machines will easily be able to read and process data published on the Web and humans will be able to use computational tools typically available in the relevant domain to work with the data.

BP13

USE LOCALE-NEUTRAL DATA REPRESENTATIONS

Humans and software agents will be able to interpret the meaning of strings representing dates, times, currencies and numbers etc. accurately.

BP14

PROVIDE DATA IN MULTIPLE FORMATS

As many users as possible will be able to use the data without first having to transform it into their preferred format.

BP15
REUSE VOCABULARIES, PREFERABLY STANDARDIZED ONES

Interoperability and consensus among data publishers and consumers will be enhanced.

BP16
CHOOSE THE RIGHT FORMALIZATION LEVEL

The most likely application cases will be supported with no more complexity than necessary.

BP17
PROVIDE BULK DOWNLOAD

Large file transfers that would require more time than a typical user would consider reasonable will be possible via dedicated file-transfer protocols.

BP18
PROVIDE SUBSETS FOR LARGE DATASETS

Humans and applications will be able to access subsets of a dataset, rather than the entire thing, with a high ratio of needed to unneeded data for the largest number of users.
Static datasets that users in the domain would consider to be too large will be downloadable in smaller pieces.
APIs will make slices or filtered subsets of the data available, the granularity depending on the needs of the domain and the demands of performance in a Web application.

BP19
USE CONTENT NEGOTIATION FOR SERVING DATA AVAILABLE IN DIFFERENT FORMATS

Content negotiation will enable different resources or different representations of the same resource to be served according to the request made by the client.

BP20
PROVIDE REAL-TIME ACCESS

Applications will be able to access time-critical data in real time or near real time, where real-time means a range from milliseconds to a few seconds after the data creation.

BP21
PROVIDE DATA UP TO DATE

Data on the Web will be updated in a timely manner so that the most recent data available online generally reflects the most recent data released via any other channel. When new data becomes available, it will be published on the Web as soon as practical thereafter.

BP22
PROVIDE AN EXPLANATION FOR DATA THAT IS NOT AVAILABLE

Consumers will know that data that is referred to from the current dataset is unavailable or only available under different conditions.

BP23

MAKE DATA AVAILABLE THROUGH AN API

Developers will have programmatic access to the data for use in their own applications, with data updated without requiring effort on the part of consumers. Web applications will be able to obtain specific data by querying a programmatic interface.

BP24

USE WEB STANDARDS AS THE FOUNDATION OF APIS

Developers who have some experience with APIs based on Web standards, such as REST, will have an initial understanding of how to use the API. The API will also be easier to maintain.

BP5

PROVIDE COMPLETE DOCUMENTATION FOR APIs

Developers will be able to obtain detailed information about each call to the API, including the parameters it takes and what it is expected to return, i.e., the whole set of information related to the API. The set of values — how to use it, notices of recent changes, contact information, and so on — should be described and easily browsable on the Web. It will also enables machines to access the API documentation in order to help developers build API client software.

BP26

AVOID CHANGES THAT MAY AFFECT THE PERFORMANCE TO YOUR API

Developer code will continue to work. Developers will know of improvements you make and be able to make use of them. Breaking changes to your API will be rare, and if they occur, developers will have sufficient time and information to adapt their code. That will enable them to avoid breakage, enhancing trust. Changes to the API will be announced on the API's documentation site.

BP27

PRESERVE IDENTIFIERS

The URI of a resource will always dereference to the resource or redirect to information about it.

BP28

ASSESS DATASET COVERAGE

Users will be able to make use of archived data well into the future.

BP29

GATHER FEEDBACK FROM DATA CONSUMERS

Data consumers will be able to provide feedback and ratings about datasets and distributions.

BP30
MAKE FEEDBACK AVAILABLE

Consumers will be able to assess the kinds of errors that affect the dataset, review other users' experiences with it, and be reassured that the publisher is actively addressing issues as needed. Consumers will also be able to determine whether other users have already provided similar feedback, saving them the trouble of submitting unnecessary bug reports and sparing the maintainers from having to deal with duplicates.

BP31
ENRICH DATA BY GENERATING NEW DATA

Datasets with missing values will be enhanced by filling in those values. Structure will be conferred and utility enhanced if relevant measures or attributes are added, but only if the addition does not distort analytical results, significance, or statistical power.

BP32
PROVIDE COMPLEMENTARY PRESENTATIONS

Complementary data presentations will enable human consumers to have immediate insight into the data by presenting it in ways that are readily understood.

BP33
PROVIDE FEEDBACK TO THE ORIGINAL PUBLISHER

Better communication will make it easier for original publishers to determine how the data they post is being used, which in turn helps them justify publishing the data. Publishers will also be made aware of steps they can take to improve their data. This leads to more and better data for everyone.

BP34
FOLLOW LICENSING TERMS

Data publishers will be able to trust that their work is being reused in accordance with their licensing requirements, which will make them more likely to continue to publish data. Reusers of data will themselves be able to properly license their derivative works.

BP35
CITE THE ORIGINAL PUBLICATION

End users will be able to assess the trustworthiness of the data they see and the efforts of the original publishers will be recognized. The chain of provenance for data on the Web will be traceable back to its original publisher.

In order to encourage publishers to implement these Best Practices for publishing data on the Web, a series of benefits were identified which may be achieved based on their implementation: comprehension, processability, discoverability, reuse, trust, data linkability, access and interoperability. These benefits are important since they provide data publishers a better understanding of "what will we possible" once the best practices are implemented. Each benefit is associated to one or more Best Practices. For example, "comprehension" is associated to ten Best Practices that are related to metadata, data vocabularies, feedback and data enhancement. This means that, if a data publisher implements these practices, the level of comprehension will increase, i.e. persons will understand better the structure and meaning of data, as well as the nature of the dataset. It is important to point out that the benefit becomes stronger as the implementation of the Best Practices increases. Taking into consideration that publishing data on the Web is an incremental process, the level of each benefit may increase after a few iterations of the data publishing process.

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.

Trust: the confidence that consumers have in the dataset will improve.

Linkability: it will be possible to create links between data resources (datasets and data items).

Access: humans and machines will be able to access up to date data in a variety of forms.

Interoperability: it will be easier to reach consensus among data publishers and consumers.

TECHNIQUES FOR PUBLISHING DATA ON THE WEB

To the extent in which the Web began consolidating itself as a platform for publishing and exchanging documents, organizations became interested in using the Web as a platform for publishing data. During the last years, different techniques have emerged for publishing data on the Web which go from the use of forms for making queries to retrieve data from a database to the publishing of Linked Data (CERI et al., 2013 and FERRARA et al., 2004). We provide below some of these techniques for publishing data (CERI et al., 2013 and FERRRAR et al., 2014), including, among these, the use of Web APIs, insertion of data directly into HTML pages and tools for the creation of data catalogs.

ACCESS USING WEB APIs

A way of publishing data on the Web consists in using Web APIs. One of the first proposals for the standardization of Web APIs were the Web Services (ALONSO et al., 2004), inspired by the RPC (Remote Procedure Call) paradigm (NELSON, 1981) and the use of XML (Extensible Markup Language) for data exchange. Subsequently, the REST (Representational State Transfer) paradigm emerged and the JSON (JavaScript Object Notation) format (MANDEL, 2008) was widely implemented. This new type of API is known as RESTful service.

In general, data exposed through APIs cannot be found by search engines. One of the reasons of the above is that, in many cases, it is necessary to do an authentication before having access to the API. Moreover, there are restrictions regarding the use of the API in order to avoid extensive access to data.

Therefore, it may be said that the data available through APIs are similar to the data available in the Deep Web, i.e. it cannot be easily found and cataloged. However, this is due to very different reason which consists in the publishers' need for controlling access to the data by external applications.

HTML PAGE ENHANCEMENT

Another way of publishing data on the Web consists in including data in HTML pages. This may be done by using microformats, i.e. specific markups (tags) that make data semantics explicit. The use of microformats makes it easier for search engines to identify data available on HTML pages and therefore show better results to users. Moreover, data publishers may achieve greater visibility. The community has developed several microformats for publishing data of different domains, including, among these: *hCalendar* for events, *hReview* for reviews and ratings, *hRecipe* for food recipes and *hCard* for personal data[12].

The use of microformats is a simple solution for publishing data on the Web; however, it may also have certain limitations: I) the use of different microformats in a same page may lead to conflicts of names (e.g. the URL class of a CSS file and the URL term of the microformat *hCalendar*), II) it does not allow making specializations and generalizations; and III) each microformat requires a specific parser.

These issues may be solved by using RDFa[13], which is a solution that allows making a specification for attributes to express structured data in any markup language, specifically XHTML[14] and HTML. While microformats combine syntax for including structured data in HTML pages with its own semantics, RDFa only cares about syntax for the inclusion of structured data. For the semantics of data, RDFa allows the use of specific vocabularies, such as schema.org[15]. RDFa allows using multiple vocabularies together with no need of specific parsers for each of them.

In addition to using RDFa for adding structured metadata in an HTML document, the JSON-LD[16] (JSON for Linked Data) language may also be used. It is a standard based on the JSON format, but it also allows the use of vocabularies and ontologies for expressing data. The JSON-LD format is highly used by the tech community and Google[17] recommends its use as standard format for exchanging Linked Data in websites.

DATA CATALOGING TOOLS

With the increasing interest for publishing open data, especially open government data, a new way of publishing data on Web stands out: data cataloging tools, such as CKAN[18] and Socrata[19]. Open data portals are created from these platforms, which provide access to previously-cataloged datasets. Datasets are organized as a series of resources and may be classified according to tags that make the data domain explicit.

Data portals are an excellent tool for cataloging datasets since they do not allow making searches in the actual datasets. In a few cases, cataloging tools offer data access APIs, although this is made in a very simplified manner. The datasets available in catalogs may be found by search engines, but it is still impossible to find elements of specific data stores in a dataset.

Despite of the great diffusion of open data portals, these solutions have different limitations, including, but not limited to: the difficulty of keeping data updated, the lack of metadata standards for the description of data sets and the impossibility of making queries about the data. Moreover, redundancy may exist since the datasets published in portals are generally available in different formats, i.e. there are multiple files for a same dataset.

[12] http://microformats.org
[13] http://w3.org/TR/rdfa-primer
[14] http://w3.org/TR/xhtml1
[15] http://schema.org

https://www.w3.org/TR/json-ld [16]
https://developers.google.com/search/docs/guides/intro-structured-data [17]
http://ckan.org [18]
http://www.socrata.com [19]

CONCLUSION

The interest for publishing data on the Web is not something new. However, the increasing interest for using the Web as a platform for sharing data brings new challenges for publishing data in a structured manner. In scenarios in which data consumers do not know each other in advance, data publishing must be done in order to satisfy groups of consumers with different requirements and profiles.

In this context, in addition to the basic aspects of the availability of data, it is also necessary to take into account other aspects related to comprehension, reliability and automatic processing of data. On one hand, publishers must provide information that helps understand the data, such as structural metadata, but must also provide information which allows consumers know the provenance and quality of the data. On the other hand, consumers must be able to provide feedback regarding the data used in order to contribute to the improvement of the publishing process. Moreover, consumers must provide information regarding the use of data, i.e. information must be provided on the used data together with the application or visualization generated based on the published data. For purposes of facilitating the tasks of publishers and consumers of data on the Web, a set of Best Practices has been proposed which addresses aspects related to the entire Life Cycle of Data on the Web. The implementation of these Best Practices leads to the creation of a communication channel between providers and consumers, in addition to improving the process for publishing data on the Web.

REFERENCES

ABITEBOUL, Serge; BUNEMAN, Peter; SUCIU, Dan. Data on the Web: from relations to semistructured data and XML. San Francisco: Morgan Kaufmann, 2000.

ALONSO, Gustavo et al. Web Services: Concepts, Architectures and Applications. Heidelberg: Springer, 2004.

BERNERS-LEE, Tim; CONNOLLY, Dan; SWICK, Ralph R. Web Architecture: Describing and Exchanging Data. 1999. Disponible en: <https://www.w3.org/1999/04/WebData>. Accedido el 04 de septiembre de 2018.

BERNERS-LEE, Tim. Linked Data. 2006. Disponible en: <https://www.w3.org/DesignIssues/LinkedData.html>. Accedido el 04 de septiembre de 2018.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, v. 5, n. 3, p.1-22, jul. 2009. IGI Global.

CERI, Stefano et al. Web Information Retrieval. Springer Science & Business Media, 2013.

CYGANIAK, Richard; WOOD, David; LANTHALER, Markus. RDF 1.1 Concepts and Abstract Syntax. 2014. Disponible en: <https://www.w3.org/TR/rdf11-concepts/>. Accedido el 04 de septiembre de 2018.

ELMASRI, Ramez; NAVATHE, Shamkant. Fundamentals of Database Systems. Addison-wesley Publishing Company, 2010.

FERRARA, Emilio et al. Web data extraction, applications and techniques: A survey. Knowledge-based Systems, [s.l.], v. 70, p.301-323, nov. 2014. Elsevier BV.

GOLDSTEIN, Brett; DYSON, Lauren (Ed.). Beyond Transparency: Open Data and the Future of Civic Innovation. San Francisco: Code for America Press, 2013.

HEATH, Tom; BIZER, Christian. Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers, 2011. 136 p. (Synthesis Lectures on the Semantic Web: Theory and Technology).

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. Dados abertos conectados. San Pablo Paulo: Novatec, 2015. 175 p.

JACOBS, Ian; WALSH, Norman. Architecture of the World Wide Web. 2004. Disponível em: <https://www.w3.org/TR/webarch/>. Accedido el 04 de septiembre de 2018.

LEE, Deirdre; LÓSCIO, Bernadette Farias; ARCHER, Phil. Data on the Web Best Practices Use Cases & Requirements. 2015. Disponible en: <https://www.w3.org/TR/dwbp-ucr/>. Accedido el 04 de septiembre de 018.

LÓSCIO, Bernadette Farias; BURLE, Caroline; CALEGARI, Newton. Data on the Web Best Practices. 2017. Disponible en: <https://www.w3.org/TR/dwbp/>. Accedido el 04 de septiembre de 2018.

MAALI, Fadi; ERICKSON, John. Data catalog vocabulary (DCAT). 2014. Disponível em: <https://www.w3.org/TR/vocab-dcat/>. Accedido el 04 de septiembre de 2018.

NELSON, Bruce Jay. Remote procedure call. 1981. 201 f. Tesis (Doctorado) - School of Computer Science, Carnegie Mellon University, Pa, 1981.

OPEN KNOWLEDGE. Open data handbook. 2012. Disponível em: <http://opendatahandbook.org/>. Acesso em: 04 set. 2018.

PIRES, Marco Túlio. Guia de Dados Abertos. São Paulo: Este Guia é parte integrante do Projeto de Cooperação entre o Governo do Estado de São Paulo e o Reino Unido, 2015. Disponible en: <http://ceweb.br/media/docs/publicacoes/13/Guia_Dados_Abertos.pdf>. Accedido el 04 de septiembre de 2018.

STRONG, Diane M.; LEE, Yang W.; WANG, Richard Y. Data quality in context. Magazine Communications of the ACM, Nueva York, v. 40, n. 5, p.103-110, 05 de mayo de 1997.

TAUBERER, Joshua; LESSIG, Larry. The 8 Principles of Open Government Data. 2007. Disponible en: <https://opengovdata.org/>. Accedido el 04 de septiembre de 2018

ATTACHMENT

ROADMAP FOR OPEN DATA PUBLISHING

1.PREPARATION

WHAT TO DO?	HOW TO DO IT?	MECHANISMS	METADATA
Identify data demands	**1.** Interact with potential consumers through interviews or public enquiries. **2.** Analyze the requests for access to information. **3.** Evaluate corporate portals or other data diffusion sources.	Data demand plan	
Identify potential datasets	**1.** Gather the demands related to similar data elements into a same dataset.	Dataset list	Descriptive
Define the priority of datasets to be opened	**1.** Define the priority of opening each dataset according to the number of requesters of each demand.	Priority list for opening	

2.CREATION

WHAT TO DO?	HOW TO DO IT?	MECHANISMS	METADATA
Dataset modeling	**1.** Evaluate the properties of each demand associated to the dataset in order to define the structure of the set as a whole. **2.** Gather similar properties, eliminate redundant properties.	Initial dataset schema	Structural
Identify the original data source	**1.** Evaluate the existing systems and documents in order to identify the original data source.	List of original data sources	Provenance
Mapping between original sources and datasets	**1.** Establish the relationship between the properties of the dataset schema and properties of the original data sources.	Mapping document between the source and the dataset	Descriptive
Identify sensitive data	**1.** Check with specialists or the applicable legislation in order to identify sensitive data.	List of sensitive data	
Identify vocabularies	**1.** Evaluate the use of known vocabularies in the definition of properties of the dataset (e.g. Dublin Core (dcterms), Friend of a Friend (FOAF), schema.org) **2.** Search in vocabulary repositories in order to identify vocabularies that are appropriate for the domain.	List of vocabularies to be used in the set schema	

2.CREATION

WHAT TO DO?	HOW TO DO IT?	MECHANISMS	METADATA
Mapping between vocabularies and dataset schema	**1.** Establish the relationship between the properties of the dataset schema and the terms of previously selected vocabularies.	Mapping document between the schema and vocabularies	
Define the data extraction strategy	**1.** According to the type of data source (e.g. database, spreadsheet, text document), specify the data extraction method.	Data extraction plan	Provenance
Define data subsets	**1.** If the volume of data is too big, define possible datasets. **2.** The division of subsets may be performed, por example, on the basis of any temporary or spatial attribute. Other more specific attributes may also be used.	List of dataset subsets	
Generate distributions	**1.** Implement a predefined extraction strategy and generate the intended data distributions.	Dataset distributions	Descriptive of distributions

4.PUBLISHING

WHAT TO DO?	HOW TO DO IT?	MECHANISMS	METADATA
Publish the dataset in a data cataloging tool	**1.** The process may vary depending on the tool to be used. In general, it is necessary to upload the files of distributions and metadata of the dataset. **2.** Complete all requested metadata and add new metadata if necessary.	Dataset available for access and download in the cataloging tool	Descriptive, versioned
Publish the dataset in an HTML page	**1.** Create the HTML page, both in a version to be consumed by persons and a machine-readable version. **2.** Insert RDFa tags in the HTML code with the semantic information for the machine-readable version.	Dataset available for access and download in an HTML page	Descriptive, versioned
Develop a data access API	**1.** Create an API which allows accessing to datasets. **2.** Create the API documentation.	Dataset available for access and download through an API and API documentation	Descriptive, versioned
Establish a communication channel with data consumers	**1.** The communication channel will depend on how the dataset was published. If the tool used does not provide a communication channel, create an HTML page.	Contact page	Use of data

3.EVALUATION

WHAT TO DO?	HOW TO DO IT?	MECHANISMS	METADATA
Evaluate the quality of data	**1.** Define the quality criteria to be evaluated (e.g. "completeness", accuracy, current nature of the data). **2.** Define metrics for th evaluation of the criteria. **3.** Define minimum requirements per each quality criterion. **4.** Evaluate the quality criteria manually or automatically.	Data quality document	Data quality
Release data for its publishing	**1.** Complete the dataset's release document.	Dataset release document	Descriptive
Return the dataset to the creation phase	**1.** Complete a document for returning the dataset to the creation phase with due justification and a description of the necessary improvements.	Document of return to creation phase