





Do the European Data Portal Datasets in the Categories Government and Public Sector, Transport and Education, Culture and Sport Meet the Data on the Web Best Practices?

Morgana Carneiro Andrade ¹, Rafaela Oliveira da Cunha ², Jorge Figueiredo ³ and Ana Alice Baptista ^{4,*}

¹ Central Library, Campus de Goiabeiras, Espírito Santo Federal University, Vitória 29075-910, Brazil; morganaandrade@hotmail.com

² Department of Information Systems, Campus de Azurém, University of Minho, 4800-058 Guimarães, Portugal; a78426@alunos.uminho.pt

³ Department of Mathematics, Campus de Gualtar, University of Minho, 4710-057 Braga, Portugal; jmfiguei@math.uminho.pt

⁴ Algoritmi Center, Campus de Azurém, University of Minho, 4800-058 Guimarães, Portugal

* Correspondence: analice@dsi.uminho.pt; Tel.: +351-253510319

Abstract: The European Data Portal is one of the worldwide initiatives that aggregates and make open data available. This is a case study with a qualitative approach that aims to determine to what extent the datasets from the Government and Public Sector, Transport, and Education, Culture and Sport categories published on the portal meet the Data on the Web Best Practices (W3C). With the datasets sorted by last modified and filtered by the ratings Excellent and Good+, we analyzed 50 different datasets from each category. The analysis revealed that the Government and Transport categories have the best-rated datasets, followed by Transportation and, lastly, Education. This analysis revealed that the Government and Transport categories have the best-rated datasets and Education the least. The most observed BPs were: BP1, BP2, BP4, BP5, BP10, BP11, BP12, BP13C, BP16, BP17, BP19, BP29, and BP34, while the least observed were: BP3, BP7H, BP7C, BP13H, BP14, BP15, BP21, BP32, and BP35. These results fill a gap in the literature on the quality of the data made available by this portal and provide insights for European data managers on which best practices are most observed and which ones need more attention.

Dataset: <https://doi.org/10.34622/datarepositorium/N2P0NK>.

Dataset License: <https://creativecommons.org/publicdomain/zero/1.0/>.

Keywords: Data on the Web Best Practices; data quality; European Data Portal; government open data



Citation: Andrade, M.C.; Cunha, R.O.d.; Figueiredo, J.; Baptista, A.A. Do the European Data Portal Datasets in the Categories Government and Public Sector, Transport and Education, Culture and Sport Meet the Data on the Web Best Practices? *Data* **2021**, *6*, 94. <https://doi.org/10.3390/data6080094>

Academic Editor: Andrea Prati

Received: 15 July 2021

Accepted: 16 August 2021

Published: 19 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Summary

The definition of data can vary remarkably between researchers and, even more so, in different knowledge domains. This diversity around the concept of data is because data are generated for various purposes, by multiple communities and processes. Data can be understood as a “(. . .) unit of content necessarily related to a certain context and composed by the triad entity, attribute and value, in such a way that, even if the details about the context of the content are not explicit, it should be implicitly available to the user, thus allowing its full interpretation” [1] (p. 2005). A dataset is a “collection of data, published or curated by a single agent, and available for access or download in one or more serializations or formats” [2], usually presented as a table [1]. Regardless of the kinds of data, they should be related to metadata, adding value to data mainly in terms of description, management, legal requirements, technical functionality, use, and preservation [3,4]. Metadata are data about data or structured data about data which, in the context of computer science and information science, are attributes that represent

the data, such as authorship, classification, description, policy, distribution terms, and copyright [1,5]. Good quality metadata help people discover and reuse datasets [6].

Currently, public sector aggregators collect large amounts of data that will later be published and made available in a single portal as open data. Open data means all "(...) information collected, produced or paid for by public bodies and can be freely used, modified and shared by anyone for any purpose" [7].

Open data are seen as an "(...) essential resource for economic growth, job creation, and societal progress" [8]. Open data bring numerous benefits, such providing insight that aids in decision making, whether in a visualized form or by reference, and they help to realize the importance of reusing data. The sector that benefits the most from open data is the public sector, indicating that the public sector is the first reuser of its data [8].

"Data portals are web-based interfaces designed to make it easier to find reusable information. (...), they contain metadata records of datasets published for reuse, mostly relating to information in the form of raw, numerical data" (...) [9].

As far as open data portals are concerned, they increasingly enable finding datasets, making possible the interaction between data publishers and reusers through forums and feedback from data and classification systems [10]. Simperl and Walker [6] (p. 16) present ten ways for open data portals to evolve to achieve sustainability and added value: "be discoverable, be measurable, promote use, organize for use, be accessible, promote standards, publish metadata, provide linkage data, co-locate documentation, and provide co-location tools". An example of a data portal is the Portuguese Open Data Portal, dados.gov.pt, or the European Data Portal (EDP), data.europa.eu.

One of the global initiatives that aggregates and give access to open data is the European Data Portal (EDP). The first version of the EDP was made available in 2016. The EDP harvests metadata available on public data and geospatial portals across European countries, which include EU member countries, EFTA countries, and countries involved in EU neighborhood policy. The datasets include, for example, land records, state maps, and the location of post offices. Access to the portal is possible through machine-readable API and human-readable websites [11,12]. In addition to this, the portal also provides thirteen data categories defined in accordance with Eurovoc domains. This thesaurus enables users to conduct multilingual searches by data categories and subject [11,13]. The EDP also aims to promote the accessibility and value of open data.

As with other initiatives, there is great concern on the part of EDP regarding data quality. In this sense, the EDP evaluates the quality of the datasets harvested concerning the FAIR principles. The FAIR principles, an acronym adopted for fairness, accuracy, interoperability, and reuse, were introduced in 2014, to "guide data producers and publishers (...) helping to maximize the added-value gained by contemporary, formal scholarly digital publishing" [14] (p. 1). The authors point out that the FAIR principles apply "(...) not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals" [14] (p. 1). The adoption of the FAIR principles enhances interoperability between different data environments [14]. Although the portal adopts a comprehensive evaluation based on the FAIR principles, some aspects are not contemplated.

On 31 January 2017, the W3C released a recommendation with 35 best practices (BPs) for publishing data on the web, named Data on the Web Best Practices (DWBP) [13]. This set of BPs addresses several challenges encountered in publishing and reusing data. The DWBP specification assigns each BP one or more benefits, out of the following eight: comprehension, processability, discoverability, reuse, trust, linkability, access, and interoperability. The following briefly presents the 35 BPs [13], as well as the benefits they can provide:

- Best Practice 1: Provide metadata—provide metadata for both human users and computer applications. This BP provides the following benefits: reuse, understanding, discovery, and processability.

- Best Practice 2: Provide descriptive metadata—the general characteristics of datasets and their distributions, facilitating their discovery on the web, as well as the nature of the datasets. Benefits: reuse, comprehension, and discoverability.
- Best Practice 3: Provide structural metadata—the schema and internal structure of distribution (e.g., description of a CSV file, an API, or an RSS feed). Benefits: reuse, comprehension, and processability.
- Best Practice 4: Provide data license information—using a link or copy of the data license agreement. Benefits: reuse and trust.
- Best Practice 5: Provide data provenance information—the origins of the data and also of all the changes they have already undergone. Benefits: reuse, comprehension, and trust.
- Best Practice 6: Provide data quality information—“provide information about data quality and fitness for particular purposes”. The quality of data should be documented and explicitly. Benefits: reuse and trust.
- Best Practice 7: Provide a version indicator—“assign and indicate a version number or date for each dataset”. Benefits: reuse and trust.
- Best Practice 8: Provide version history—making a description for each version available that explains how it differs from the previous version. Benefits: reuse and trust.
- Best Practice 9: Use persistent URIs as identifiers of datasets—enables the identification of datasets in a persistent way. Benefits: reuse, interoperability, and linkability.
- Best Practice 10: Use persistent URIs as identifiers within datasets—reuse URIs between datasets and ensure that their identifiers can be referred to by other datasets consistently. Benefits: reuse, interoperability, linkability, and discoverability.
- Best Practice 11: Assign URIs to dataset versions and series—to individual versions of datasets, as well as to the overall series. Benefits: reuse, discoverability, and trust.
- Best Practice 12: Use machine-readable standardized data formats—to minimize the limitations on the use of data. Benefits: reuse and processability.
- Best Practice 13: Use locale-neutral data representations—to limit misinterpretations; if this is not possible, metadata on the locality used by the data values must be provided. Benefits: reuse and comprehension.
- Best Practice 14: Provide data in multiple formats—to reduce costs in transforming datasets and mistakes during the process. Benefits: reuse and processability.
- Best Practice 15: Reuse vocabularies, preferably standardized ones—to encode data and metadata. Benefits: reuse, processability, comprehension, trust, and interoperability.
- Best Practice 16: Choose the right formalization level—the level that fits the most likely data and applications. Benefits: reuse, comprehension, and interoperability.
- Best Practice 17: Provide bulk download—in a way that allows consumers to retrieve the complete dataset with a single request. Benefits: reuse and access.
- Best Practice 18: Provide subsets for large datasets—so that data users can download only the subset they need. Benefits: reuse, linkability, access, and processability.
- Best Practice 19: Use content negotiation for serving data available in multiple formats—to serve data available in various formats. Benefits: reuse and access.
- Best Practice 20: Provide real-time access—for immediate access to encourage the development of real-time applications. “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”. Benefits: reuse and access.
- Best Practice 21: Provide data that is up to date—and make the frequency of updating explicit. Benefits: reuse and access.
- Best Practice 22: Provide an explanation for data that is not available—“provide an explanation of how the data can be accessed and who can access it”, to provide full context for potential data consumers. Benefits: reuse and trust.

- Best Practice 23: Make data available through an API—to offer the greatest flexibility and processability for the data consumers. Benefits: reuse, processability, interoperability, and access.
- Best Practice 24: Use web standards as the foundation of APIs—so that they are more usable and leverage the strengths of the web. APIs should be built on web standards to leverage the strengths of the web (e.g., REST). Benefits: reuse, processability, access, discoverability, and linkability.
- Best Practice 25: Provide complete documentation for your API—in a way that developers perceive its quality and usefulness. “Update documentation as you add features or make changes”. Benefits: reuse and trust.
- Best Practice 26: Avoid breaking changes to your API—so that the client code does not stop working. Benefits: trust and interoperability.
- Best Practice 27: Preserve identifiers—if it is necessary to remove the data from the web, it is necessary to preserve the respective identifiers so that the user is not directed to the 404 response code (not found). Benefits: reuse and trust.
- Best Practice 28: Assess dataset coverage—assess the coverage of a dataset before its preservation. Benefits: reuse and trust.
- Best Practice 29: Gather feedback from data consumers—through an easily detectable mechanism. “Data consumers will be able to provide feedback and ratings about datasets and distributions”. Benefits: reuse, trust, and comprehension.
- Best Practice 30: Make feedback available—give publicly available consumer feedback about datasets and distributions datasets. Benefits: reuse and trust.
- Best Practice 31: Enrich data by generating new data—to enhance their value. Benefits: reuse, comprehension, trust, and processability.
- Best Practice 32: Provide complementary presentations—such as visualizations, tables, web applications, and summaries. Benefits: reuse, comprehension, access, and trust.
- Best Practice 33: Provide feedback to the original publisher—on, for example, when and how their data are being reused or aspects of improvement. Benefits: reuse, interoperability, and trust.
- Best Practice 34: Follow licensing terms—in order to maintain a good relationship with the original publisher. Benefits: reuse and trust.
- Best Practice 35: Cite the original publication—in order to generate trust in the data. Benefits: reuse, trust, and discoverability.

In this study, we try to determine to what extent the datasets from the Government and Public Sector, Transport, and Education, Culture and Sport categories published on the portal meet the Data on the Web Best Practices (W3C).

2. Data Description

This section presents the data resulting from the study, whose methodology is described in Section 3 below.

A total of 150 datasets were analyzed in light of 29 BPs and, because some were targeted to both humans and machines, a total of 4350 analyses were performed.

The number of datasets observing or not observing each BP in the Government and Public Sector category is presented in Table 1 and Figure 1.

Table 1. Number of datasets in the Government and Public Sector category observing or not observing each BP.

Best Practices	Observed by	Not Observed by
BP1	50	
BP2	38	12
BP3		50
BP4	50	
BP5	49	1
BP6		50
BP7H, the human-readable version of BP7	5	45
BP7C, the computer-readable version of BP7	5	45
BP8H, the human-readable version of BP8		50
BP8C, the computer-readable version of BP8		50
BP9		50
BP10	48	2
BP11	49	1
BP12	49	1
BP13H, the human-readable version of BP13	17	33
BP13C, the computer-readable version of BP13	37	13
BP14	19	31
BP15	18	32
BP16	49	1
BP17	50	
BP19	50	
BP21	23	27
BP22H, the human-readable version of BP22		50
BP22C, the computer-readable version of BP22		50
BP29	42	8
BP30		50
BP32	23	27
BP34	50	
BP35	7	43

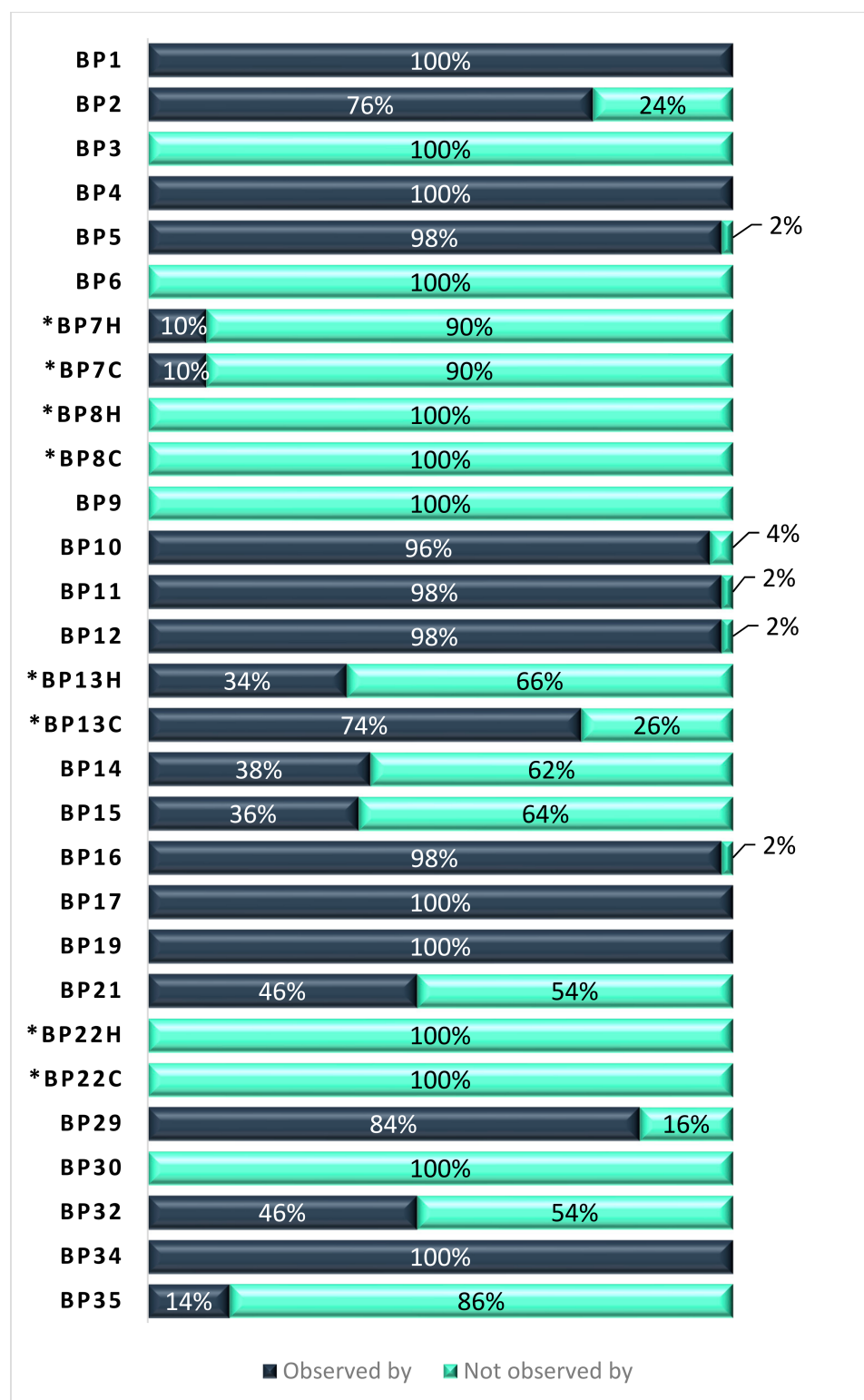


Figure 1. Percentage of datasets from the Government and Public Sector category observing or not observing each BP. Note: the figure does not show the BPs that are not applied. * BP7H, the human-readable version of BP7; BP7C, the computer-readable version of BP7; BP8H, the human-readable version of BP8; BP8C, the computer-readable version of BP8; BP13H, the human-readable version of BP13; BP13C, the computer-readable version of BP13; BP22H, the human-readable version of BP22; BP22C, the computer-readable version of BP22.

The results for the Transport category are presented in Table 2 and Figure 2.

Table 2. Number of datasets in the Transport category observing or not observing each BP.

Best Practices	Observed by	Not Observed by
BP1	50	
BP2	45	5
BP3	9	41
BP4	50	
BP5	50	
BP6		50
BP7H, the human-readable version of BP7	16	34
BP7C, the computer-readable version of BP7	16	34
BP8H, the human-readable version of BP8		50
BP8C, the computer-readable version of BP8		50
BP9		50
BP10	50	
BP11	50	
BP12	50	
BP13H, the human-readable version of BP13	9	41
BP13C, the computer-readable version of BP13	36	14
BP14	21	29
BP15	2	48
BP16	50	
BP17	50	
BP19	50	
BP21	10	40
BP22H, the human-readable version of BP22		50
BP22C, the computer-readable version of BP22		50
BP29	44	6
BP30		50
BP32	33	17
BP34	50	
BP35	4	46

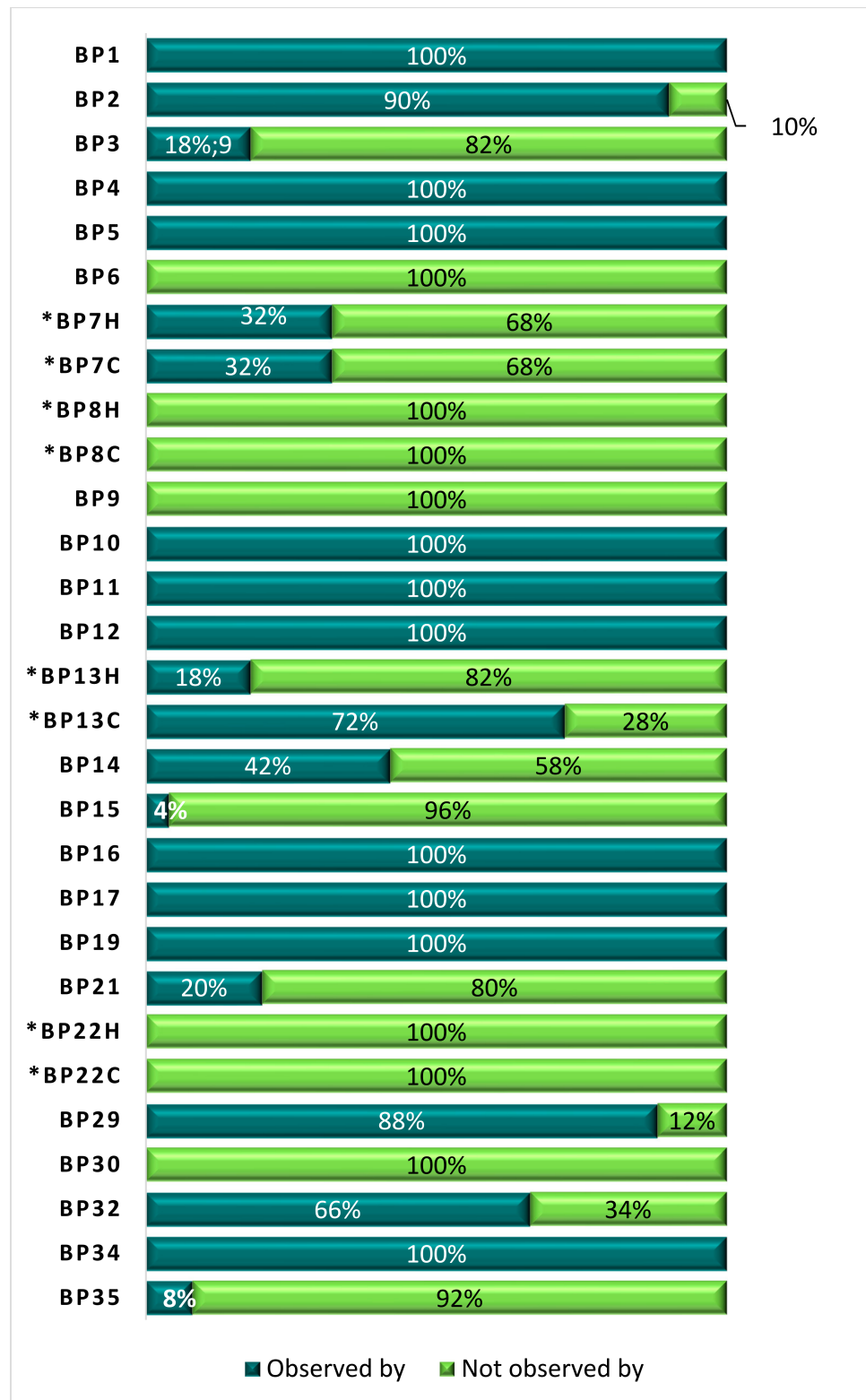


Figure 2. Percentage of datasets from the Transport category observing or not observing each BP. Note: the figure does not show the BPs that are not applied. * BP7H, the human-readable version of BP7; BP7C, the computer-readable version of BP7; BP8H, the human-readable version of BP8; BP8C, the computer-readable version of BP8; BP13H, the human-readable version of BP13; BP13C, the computer-readable version of BP13; BP22H, the human-readable version of BP22; BP22C, the computer-readable version of BP22.

The results of the Education, Culture and Sport category are displayed in Table 3 and Figure 3.

Table 3. Number of datasets in the Education, Culture and Sport category observing or not observing each BP.

Best Practices	Observed by	Not Observed by
BP1	50	
BP2	46	4
BP3	4	46
BP4	50	
BP5	50	
BP6		50
BP7H, the human-readable version of BP7	5	45
BP7C, the computer-readable version of BP7	5	45
BP8H, the human-readable version of BP8		50
BP8C, the computer-readable version of BP8		50
BP9		50
BP10	49	1
BP11	50	
BP12	48	2
BP13H, the human-readable version of BP13	14	36
BP13C, the computer-readable version of BP13	31	19
BP14	22	28
BP15	1	49
BP16	50	
BP17	50	
BP19	50	
BP21	17	33
BP22H, the human-readable version of BP22		50
BP22C, the computer-readable version of BP22		50
BP29	33	17
BP30		50
BP32	24	26
BP34	50	
BP35	7	43

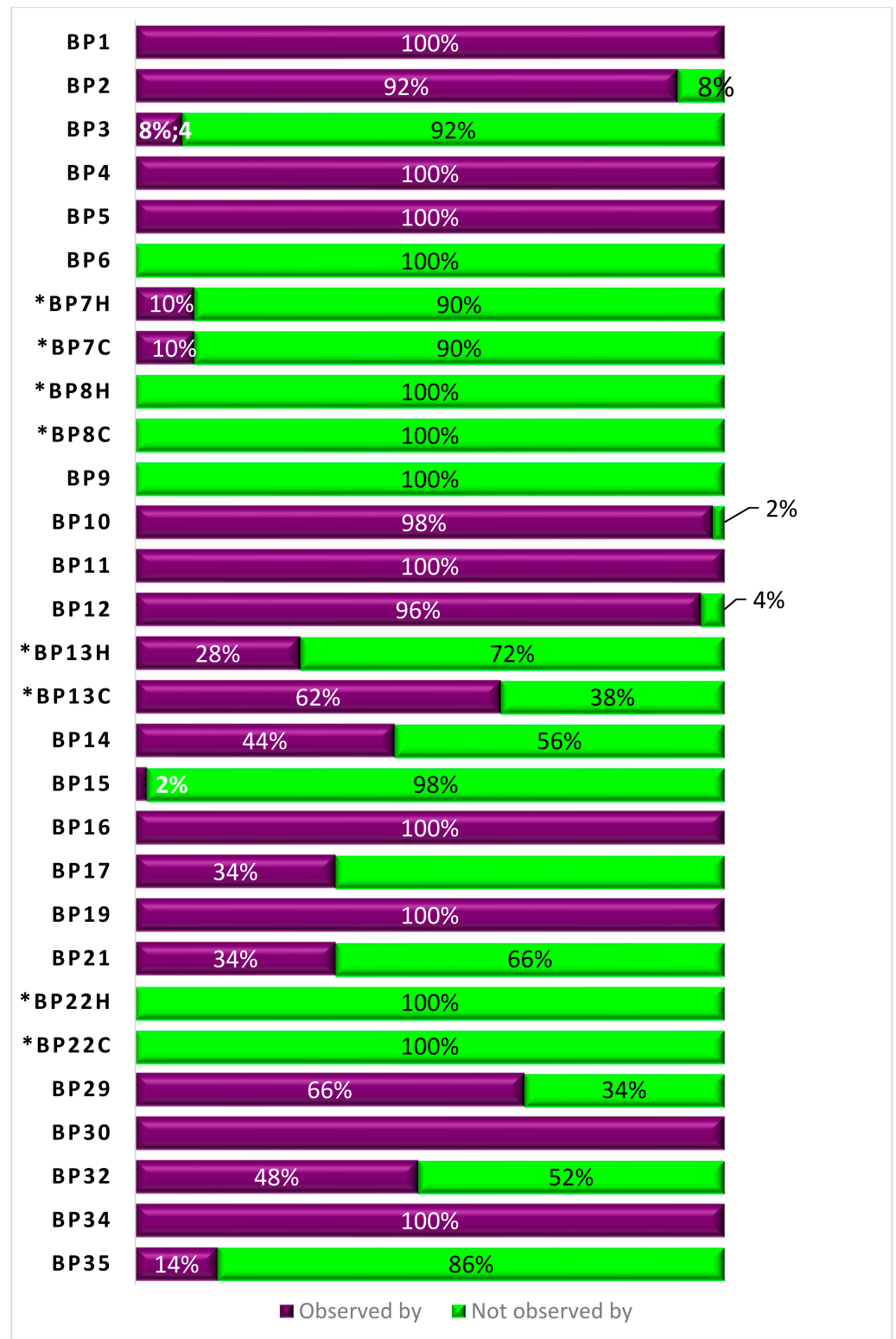


Figure 3. Percentage of datasets from the Education, Culture and Sport category observing or not observing each BP. Note: the figure does not show the BPs that are not applied. * BP7H, the human-readable version of BP7; BP7C, the computer-readable version of BP7; BP8H, the human-readable version of BP8; BP8C, the computer-readable version of BP8; BP13H, the human-readable version of BP13; BP13C, the computer-readable version of BP13; BP22H, the human-readable version of BP22; BP22C, the computer-readable version of BP22.

3. Methods

Three categories were randomly selected from the European Data Portal: Government and Public Sector, Transport, and Education, Culture and Sport. The research was conducted in two stages: an exploratory study and a final study. We prepared a spreadsheet with the BPs in rows and the datasets' identifiers in columns for both studies. The final analysis was performed and recorded by putting in each cell one of the following codes: "Yes" (Y), "No" (N). In addition, rows were added in final study sheets to 4 BPs to indicate whether those BPs correspond to machine- or human-readable data. Some BPs were not analyzed as they were out of scope for this study. In these cases, the respective cells have the value "Does Not Apply" (NA) (we provide the datasets as Supplementary Materials, Tables S1–S4, at DataRepositoriUM, <https://doi.org/10.34622/datarepositorium/N2P0NK> (accessed on 6 August 2021)).

In the final study, in addition to the best practices field, an observation row was inserted for each BP to include some notes as needed (we provide the datasets as Supplementary Materials, Tables S2–S4, at DataRepositoriUM, <https://doi.org/10.34622/datarepositorium/N2P0NK> (accessed on 6 August 2021)).

The procedures for each study are described below.

3.1. Exploratory Study

To analyze the data quality of the European Data Portal in the categories of Government and Public Sector, Transport, and Education, Culture and Sport, an exploratory study was conducted, where the first 20 datasets from each category were analyzed.

Since it was not possible to analyze all datasets manually in a timely manner, a sample had to be defined. Initially, a systematic sampling of the datasets classified as Excellent and Good+ was carried out. For this purpose, the Algorithm 1 was used.

Algorithm 1 Algorithm for datasets selection—Exploratory study

```

START
  FILTER datasets by Excellent to a new list;
  FILTER datasets by Good+ and ADD them to the list;
  count = 0;
  WHILE (count < 20)
    ADD dataset to the sample;
    REMOVE dataset from the list;
    count = count + 1;
  END_WHILE
END

```

The exploratory study only focused on the first 20 datasets of each category, as this study aimed to verify the suitability of the algorithm for constituting the sample, obtain the first results, and identify potential implementation problems. The first 20 datasets served as a test for the algorithm, and if it proved to be effective and did not limit the analysis, it would be adopted in the final study and the sample would be extended to 50 datasets of each category (we provide the dataset as Supplementary Materials, Table S1, at DataRepositoriUM, <https://doi.org/10.34622/datarepositorium/N2P0NK> (accessed on 6 August 2021)).

With the preliminary study, we verified that the sampling algorithm was not optimal, since many datasets belonging to the same country were selected, many of them similar, only differing in the modification date and the data themselves. To overcome this problem and increase the variety of datasets, some changes were made to the sampling procedure as shown in the next section.

3.2. Final Study

In the final study, the Algorithm 2 was used.

Algorithm 2 Algorithm for datasets selection—Final study

```

START
  FILTER datasets by Excellent to a new list;
  FILTER datasets by Good+ and ADD them to the list;
  REMOVE datasets from the exploratory study from the list;
  count = 0;
  WHILE (count < 50)
    IF new dataset is very similar to dataset already included in the sample THEN
      REMOVE dataset from the list;
    ELSE
      ADD dataset to the sample;
      REMOVE dataset from the list;
      count = count + 1;
    END_IF
  END_WHILE
END

```

For each category, the selection of 50 datasets was carried out based on the aforementioned algorithm, in which the datasets were filtered by the classification of Excellent or Good+ and the datasets of the exploratory study were removed. Thus, no datasets classified as Excellent were left in any category. The list was scrolled through dataset by dataset to constitute the sample with 50 datasets, according to the requirement of not including similarities.

To perform the analysis of each dataset, the human- and machine-readable (Turtle) EDP catalog information of each dataset was used. For some BPs, this study was carried out in two or more rounds because it was necessary to reverify or fine-tune information. Additionally, the following BPs were left out of the analysis as they were not applicable in the context of the EDP or in the context of this study: B18, B20, BP23, BP24, BP25, B26, B27, B28, BP31, and B33.

Although the W3C specification on DWBP is clear on how to identify compliance or non-compliance with each BP, it was necessary to further elaborate on these criteria to eliminate subjectivities and deviations in analysis over time. To facilitate the analysis and the presentation of the data, we divided some BPs into human-readable and computer-readable versions. The criteria applied in the analysis of the observation of each BP were the following:

Best Practice 1—Provide metadata. Provide metadata for human users and computer applications, so that humans can analyze the metadata and computer applications can process it. Always add the code “yes” since the European Data Portal always requires the provision of metadata.

Best practice 2—Provide descriptive metadata. If the dataset catalog provides information about the date, keywords, title, and publisher, among others, then put “yes”, otherwise put “no”. Add the unavailable metadata elements, considered essential, into the respective remarks field.

Best practice 3—Provide structural metadata. They are needed to open the dataset. If they have information about the meaning and acceptable values for each field, put “yes”, otherwise put “no”.

Best Practice 4—Provide data license information. If the dataset has a license and the license type is clear [15], then put “yes”, but add relevant information in the remarks field. If it does not have a license, then put “no”.

Best practice 5—Provide data provenance information. There is provenance information if there is information about the publisher ID, dates of creation, and modification of the dataset [16,17]. They were analyzed from two perspectives: (a) besides the dct:issued property being present, one of the following properties or all three should be present: dct:creator, dct:publisher; dct: publisher and (b) if none of the previous properties is

present, `prov:actedOnBehalfOf` should be present. If one of the two possibilities or both are present, put “yes”, otherwise put “no” and add information in an appropriate remarks field.

Best Practice 6—Provide data quality information. If the dataset has the property `dqv:hasQualityMeasurement`, put “yes”; otherwise, put “no”.

Best Practice 7—Provide a version indicator. Divided by us into BP7H, human-readable information, and BP7C, computer-readable information. In BP7H, if the dataset has version information, put “yes”. In BP7C, “yes” is only added if it has some property such as `pav:version` or `owl:versionInfo`. These properties can be identified in the Turtle syntax. In this case, appropriate information is added to the remarks field.

Best Practice 8—Provide version history. Divided by us into BP8H, human-readable information, and BP8C, computer-readable information. In BP8C put “yes” only if it has the metadata elements `dct:isVersionOf`, `dct:hasVersion`, `owl:versionInfo`, `pav:version`, or an equivalent associated with `rdfs:comment`. In this case, appropriate information is added to the remarks field. For BP8H, if a summary of the differences between versions is provided, put “yes”; otherwise, put “no”.

Best Practice 9—Use persistent URIs as identifiers of datasets. If the dataset uses known persistent identifiers such as URN, Handle, DOI, ARK, Persistent Uniform Resource Locators (PURLs), Electronic Identifier Serial Publications (EISPs), International eBook Identifier Numbers (IEINs), Extensible Resource Identifiers (XRIs), Magnetic Links—magnet, Virtual International Authority File (VIAF), International Standard Name Identifier (ISNI), or International Standard Name Identifier (ISNI), put “yes”; if not, put “no”.

Best Practice 10—Use persistent URIs as identifiers within datasets. Check if properties such as `dct:creator`, `dct:publisher`, `dct:location`, `dct:spacial`, `dct:subject`, `dct:licence`, or `dct:contributor` are referenced by a persistent URI, e.g., DOI or Handle for documents, Orcid for the author, or URI for Creative Commons license. If yes, put “yes”; otherwise put “no”.

Best practice 11—Assign URIs to dataset versions and series. If a URI is assigned to each version, put “yes”; otherwise, put “no”.

Best Practice 12—Use machine-readable standardized data formats. If the dataset has standardized machine-readable distributions, such as XML, JSON, Turtle, and/or CSV, put “yes”; otherwise, put “no”.

Best practice 13—Use locale-neutral data representations. Divided by us into BP13H, human-readable information, and BP13C, computer-readable information. In BP13H, if there is information on how to interpret the respective values in the columns (dates, times, currencies, and numbers), then put “yes”; otherwise, put “no”. For the BP13C, it was necessary to search by properties: `dct:conformsTo`, `dct:language`, `dct:location` and/or `dct:spacial`. If identified, put “yes”, otherwise, put “no”.

Best practice 14—Provide data in multiple formats. If the dataset has distributions in several formats, put “yes”; otherwise, put “no”.

Best practice 15—Reuse vocabularies, preferably standardized ones. The EDP uses DCAT and the data theme authority table, adopted for dataset classification (`dcat:theme`), by default for all datasets. Therefore, our analysis focused only on the use of value vocabularies such as Eurovoc. Thus, if the dataset descriptions has the property `dct:subject` with values from standard vocabularies (Eurovoc), put “yes”; otherwise, put “no”.

Best practice 16—Choose the right formalization level. For this best practice, if the dataset uses appropriate vocabulary, as Dublin Core and Schema.org, to describe, put “yes”; if it uses vocabulary that is not over- or underspecified, put “no”.

Best practice 17—Provide bulk download. If the dataset can be downloaded all at once, put “yes”; otherwise, put “no”.

Best Practice 18—Provide subsets for large datasets. This best practice only applies to large datasets. In the EDP, these are already divided, so it does not apply.

Best Practice 19—Use content negotiation for serving data available in multiple formats. Check the available representations of the resource and try to get them by specifying

the accepted content in the HTTP request header. If it returns, put “yes”; otherwise put “no”.

Best Practice 20—Provide real-time access. The EDP encourages data providers to make data available in real time. However, this BP cannot be verified by analysis of its catalog and was therefore not included in the analysis.

Best Practice 21—Provide data up to date. If there is the property `dct:accrualPeriodicity` or similar, put “yes”, otherwise, put “no”.

Best practice 22—Provide an explanation for data that are not available. Divided by us into BP22H, human-readable information, and BP22C, computer-readable information. For BP22H, if datasets are accompanied by an HTML document with information about data referred to in the dataset but not available for some reason, put “yes”; otherwise put “no”. For BP22C, if appropriate HTTP status codes are used, such as 303 (see others), 410 (permanently removed), or 503 (service *provides data* not available), put “yes”; otherwise, put “no”.

Best practice 23—Make data available through an API. The EDP enables the distribution of datasets by API, but as compliance with this BP does not depend on the datasets, it was not analyzed.

Best practice 24—Use web standards as the foundation of APIs. The compliance with this BP does not depend on the datasets, so it was not analyzed.

Best practice 25—Provide complete documentation for your API. The compliance with this BP does not depend on the datasets, so it was not analyzed.

Best Practice 26—Provide complete documentation for your API. The EDP provides complete documentation for their APIs. The compliance with this BP does not depend on the datasets, so it was not analyzed.

Best practice 27—Preserve identifiers. This BP is also not applicable to the scope of this study since we did not look at removed datasets.

Best Practice 28—Assess dataset coverage. The analysis of the compliance of this BP is also out of the scope of this study since it is related to preservation information for archival purposes.

Best practice 29—Gather feedback from data consumers. Data consumers will be able to provide feedback and evaluations on the datasets and their distributions. If there is a feedback mechanism for data consumers, such as email or another communication channel, put “yes”; otherwise, put “no”.

Best Practice 30—Make feedback available. The feedback may be made available to data consumers. The existence of the property was verified as `rdfs:comment` or similar and, if so, put “yes”, otherwise, put “no”.

Best Practice 31—Enrich data by generating new data. As we found no information either within the datasets or in the metadata that would allow us to say that the data were enriched, this best practice was not included in the analysis.

Best practice 32—Provide complementary presentations. If there are complementary presentations of the dataset such as a graph, put “yes”; otherwise, put “no”.

Best Practice 33—Provide feedback to the original publisher. Compliance with this BP is out of scope of this study as we had no access to the communication between the EDP and its data providers.

Best practice 34—Follow the licensing terms. Although the EDP collects data with the same type of license provided at the source (<https://data.europa.eu/pt/faq> (accessed on 28 May 2021)), it was checked whether the dataset follows the license of the data according to the presented term. If it does, put “yes”; otherwise put “no”.

Best practice 35—Cite the original publication. If the citation of the original source of any dataset was available by a text or a link (e.g., data source, available from) put “yes”; otherwise put “no”.

This analysis revealed that the Government and Transport categories have the best-rated datasets and Education the least. The most observed BPs were: BP1, BP2, BP4, BP5, BP10, BP11, BP12, BP13C, BP16, BP17, BP19, BP29, and BP34, while the least observed

were: BP3, BP7H, BP7C, BP13H, BP14, BP15, BP21, BP32, and BP35. Additionally, in the 3 categories analyzed, BP6, BP8H, BP8C, BP9, BP22H, BP22C, and BP30 were not observed by any dataset (we provide the datasets as Supplementary Materials, Tables S2–S4, at DataRepositoriUM, <https://doi.org/10.34622/datarepositorium/N2P0NK> (accessed on 6 August 2021)).

These results highlight the importance of quality-driven data publishing. Data publishing provides benefits for both managers and users. Data publication can be very useful for various sectors and users, as in the case of transport, to provide a more efficient response in emergencies [18] or by providing subsidies in decision making. However, it does not make sense to publish data without the attention that should be given to quality as it is necessary to ensure reliability in access and reuse. As well as the FAIR principles, the observance of the best practices recommended by W3C enhances the quality of open data, with DWBP being more comprehensive.

The result of this study offers insights to data managers, notably in the context of government, on which best practices are most observed and which need more attention. In addition, it fills a gap in the literature on the quality of data provided by the EDP from the DWBP perspective.

The limitation of the study was that we did not analyze BP18, BP20, BP23, BP24, BP25, BP26, BP27, BP28, BP31, and BP33 due to not meeting the scope of this study.

Despite the extra care with the sampling technique, many datasets are still similar, so new studies will need to start by refining the sample constitution algorithm.

4. User Notes

Our datasets are made available as CSV files, an open format. On the first page of each CSV, there is structural information about the data. Legends for the abbreviations are at the bottom of the CSV.

The CSV sheets are structured as follows: rows—BPs; columns—identifiers for each dataset. In the final study sheets, we included a row for remarks on each BP.

Supplementary Materials: The spreadsheets are available at <https://doi.org/10.34622/datarepositorium/N2P0NK>. Table S1: Exploratory study, Government, Transport and Education Culture Sports categories. Table S2: Final study, Government category. Table S3: Final study, Transport category. Table S4: Final study, Education, Culture and Sports.

Author Contributions: Conceptualization, A.A.B.; methodology, A.A.B. and J.F.; investigation, writing—original draft preparation, R.O.d.C. and M.C.A.; writing—review and editing, R.O.d.C., A.A.B. and M.C.A.; supervision—M.C.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This work resulted in the creation of datasets which is available at <https://doi.org/10.34622/datarepositorium/N2P0NK>, as per the CC0—“Public Domain Dedication” License, accessed on 6 August 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Santos, P.L.V.A.C.; Sant’Ana, R.C.G. Dado e granularidade na perspectiva da informação e tecnologia: Uma interpretação pela ciência da informação. *Ciênc. Inf.* **2013**, *42*, 199–209.
2. Albertoni, R.; Cox, S.; Beltran, A.G.; Prego, A.; Winstanley, P. Data Catalog Vocabulary (DCAT—Version 2. W3C Recommendation 4 February 2020). Available online: <https://www.w3.org/TR/vocab-dcat/> (accessed on 18 February 2021).
3. Greenberg, J. Metadata and the World Wide Web. *Encyclopedia of Library and Information Science*. 2003. Available online: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.4528&rep=rep1&type=pdf> (accessed on 28 May 2021).

4. Riley, J. *What Is Metadata, and What Is It for?* NISO: Baltimore, MD, USA, 2017. Available online: https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf (accessed on 28 May 2021).
5. Lee-Berners, T. Weaving the Web: Glossary. 23 July 1999. Available online: <https://www.w3.org/People/Berners-Lee/Weaving/glossary.html> (accessed on 28 May 2021).
6. Simperl, E.; Walker, J. *Analytical Report 8: The Future of Open Data Portals*; Publications Office of the European Union: Luxembourg, 2017; pp. 1–26. Available online: https://www.europeandataportal.eu/sites/default/files/edp_analyticalreport_n8.pdf (accessed on 5 October 2020).
7. Carrara, W.; Fischer, S.; van Steenberg, E. Open Data Maturity in Europe 2015: Insights into the European State of Play. 2020. Available online: <https://beta.op.europa.eu/en/publication-detail/-/publication/0e95f3cb-141c-11eb-b57e-01aa75ed71a1> (accessed on 13 October 2020).
8. Berends, J.; Carrara, W.; Radu, C. *Analytical Report 9: The Economic Benefits of Open Data*; Publications Office of the European Union: Luxembourg, 2017. [CrossRef]
9. European Commission. Open Data Portals. 2021. Available online: <https://digital-strategy.ec.europa.eu/en/policies/open-data-portals> (accessed on 28 May 2021).
10. Van Knippenberg, L. *Analytical Report 16: Open Data Best Practices in Europe: Learning from Cyprus, France, and Ireland*; Publications Office of the European Union: Luxembourg, 2020. [CrossRef]
11. Data.Europa.Eu. About Data.Europa.Eu. (2021 Update). Available online: <https://data.europa.eu/de/highlights/open-regions-and-cities-data-european-data-portal> (accessed on 21 June 2021).
12. National Spatial Data Infrastructure. About nipp.hr. 2021. Available online: <https://www.nipp.hr/default.aspx?id=1728>. (accessed on 19 June 2021).
13. Lóscio, B.F.; Burle, C.; Calegari, N. (Eds.) Data on the Web Best Practices. 31 January 2017. Available online: <https://www.w3.org/TR/dwbp/#intro>. (accessed on 13 October 2020).
14. Wilkinson, M.D.; Dumontier, M.; Jan Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.; Santos, L.B.D.; Bourne, P.E.; et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 60018. [CrossRef]
15. Torino, E.; Trevisan, G.L.; Vidotti, S.A.B.G. Dados abertos CAPES: Um Olhar à Luz dos Desafios para Publicação de Dados na Web. *Ciênc. Inf.* **2019**, *48*, 38–46. Available online: <https://repositorio.utfpr.edu.br/jspui/handle/1/4812>. (accessed on 1 June 2021).
16. Provenance. Linked Data Glossary. 2013. Available online: <https://www.w3.org/TR/ld-glossary/#provenance>. (accessed on 5 May 2021).
17. Hartig, O. Provenance Information in the Web of Data. 2009. Available online: http://ceur-ws.org/Vol-538/ldow2009_paper18.pdf (accessed on 28 May 2021).
18. Carrara, W.; Fischer, S.; Oudkerk, F.; van Steenberg, E.; Tinholt, D. *Analytical Report 1: Digital Transformation and Open Data*; Publications Office of the European Union: Luxembourg, 2015; pp. 1–22. [CrossRef]