An Assessment of Adoption and Quality of Linked Data in European Open Government Data [⋆]

Luis-Daniel Ibáñez[1][0000−0001−6993−0001], Ian Millard[2], Hugh Glaser[2], and
Elena Simperl[1][0000−0003−1722−947X]

[1] University of Southampton
{l.d.ibanez,e.simperl}@soton.ac.uk
[2] Seme4 Ltd.
{ian.millard,hugh.glaser}@seme4.com

Abstract. The European Commission has adopted Linked Data principles and practices with the purpose of increasing the accessibility, interoperability and value of the data that is made available openly by European public sector organisations. This includes investment in metadata development for describing open datasets, catalogs of resources with persistent URIs, and the European Data Portal (EDP), which provides a single point of access, search and exploration of European open data. As the Public Sector Initiative (PSI) Directive is being revised, a critical question for the Commission is the extent to which open government data publishers have adopted Linked Data, and how they are applying the underlying technologies. In this paper, we undertake a quantitative analysis to support this. We explore if and how open data portals indexed by the EDP are using Linked Data and assess the quality of the datasets according to multiple dimensions.

Keywords: Linked Data · open government data · data quality

1 Introduction

Linked Data refers to a set of principles, technologies and practices that facilitate data integration. Publishers are encouraged to adopt them to make their data more useful [5]. Linked Data makes it easier for developers to access and combine datasets from different sources. To unlock the value of their data, publishers are advised to [5]:

1. use URIs to name things and relationships among things;
2. use HTTP URIs so those names can be looked up (a technique called *dereferencing*);
3. return useful information upon lookup of URIs, using open standards such as RDF; and

[⋆] Supported by the European Data Portal, an initiative funded by the European Union

4. include links to other URIs, so more things and relationships can be discovered organically.

Public sector organisations have embraced open data as a way to increase transparency and accountability of government services, boost innovation and foster participation [2]. For this purpose, they have set up so-called *open data portals*, which are web repositories where the data released by different government agencies can be searched, explored and downloaded. Open Data Soft, a technology provider in this space, estimates that there are more than 2600 such portals around the world.[3] To help track progress in open data publishing, Sir Tim Berners-Lee developed a 5-star deployment scheme, which features Linked Data as ultimate goal:[4]

1. Publish data under an open license.
2. Publish structured data.
3. Publish data using open formats.
4. Use URIs to denote things (matching Linked Data principles 1 to 3).
5. Link data to other data to provide context (matching principle 4).

Semantic Web (SW) technologies were chosen by the EC as the vehicle to achieve seamless and meaningful cross-border and cross-domain data exchanges between public administrations. Compared to other integration technologies, SW principles ensure data exchanged between public administrations is automatically recognised thanks to unambiguous, shared meaning, and setting the field for progressive and focused data integration among member states. Linked Data and the 5-star scheme are at the core of the open data strategy of the European Commission (EC), described in the Public Sector Information (PSI) Directive. This includes investment in the development and promotion of: metadata specifications such as DCAT-AP[5] to describe datasets; catalogs of resources with persistent URIs;[6], a data portal to host EC data;[7] as well as the *European Data Portal (EDP)*,[8] which provides a single point of access, search and exploration of open government data by various European public institutions. In November 2015, the EDP has started to harvest metadata from all national portals of the 28 EU countries and associated countries, the EC data portal, and a set of other sources such as geospatial portals. As the PSI Directive is being revised, policy makers need an overview of the adoption of their original recommendations by publishers, along the following lines:

(i) Are publishers using Linked Data? (ii) Are they using RDF or do they prefer other structured formats? (iii) Is the Linked Data they generate of enough quality to be queried and re-used?

[3] https://opendatainception.io/
[4] https://5stardata.info/en/
[5] https://joinup.ec.europa.eu/release/dcat-ap/12
[6] http://data.europa.eu/URI.html
[7] http://data.europa.eu/euodp/en/home
[8] https://www.europeandataportal.eu/

In this paper, we present a quantitative study that helps answer these questions. We analyse the use of Linked Data and the extent to which publishers indexed by the EDP follow the core principles. We explore the following themes: (i) the uptake of RDF as a publishing format, compared to structured and unstructured alternatives; (ii) the quality of the linked datasets, as an indicator of how well publishers implement Berners-Lee's deployment scheme; item and in comparison with previous quality assessments of the general *linked open data (LOD) cloud*.

Our contribution to the semantic web community is an up-to-date, empirically grounded reality check of the acceptance and uptake of arguably one of its core achievements - the principles, technologies and practices around Linked Data - in a critical early adopter sector, using a representative sample which includes 78 data portals, including all EU countries. We offer insight into how government publishers go about producing Linked Data and identify challenges and areas of improvement, which should inform the design of new supporting tools and techniques.

2 Related Work

The public sector has been one of the supporters of Linked Data from the beginning. There is a large body of literature documenting major open government data projects in different countries [6, 15], compiling methodological guidance [16], and providing technical support [10, 11, 3].

Several studies have focused on empirical data quality assessments of different snapshots of the linked open data web. For instance, [7] looked at a corpus of more than one billion quadruples from almost four million documents acquired in 2010. [14] analysed best practice adoption in terms of linking, vocabulary usage, and metadata provision in different topical domains from a sample of 1014 datasets (including 183 from government) crawled in 2014. [4] evaluated the quality of a crawl seeded from the LOD cloud 2014 dataset of 130 datasets, totalling approximately 3.7 Billion quads. Our study applies a subset of the metrics used in these previous works on a much more recent (March 2019) corpus of open government datasets. Our study is also novel in the sense of analysing the unique perspective of a metaportal like the EDP, that by design is constrained to certain publishers and their catalogs.

Initiatives such as Open Data Monitor[9] and Portal Watch[10] keep track of a sample of web-based data portals and evaluate them according to criteria such as availability, conformance, retrievability, accuracy and metadata openness [13]. Our work complements them with a focused analysis of the quality of the datasets published as Linked Data on open government portals.

[9] https://opendatamonitor.eu/

[10] https://data.wu.ac.at/portalwatch/

3 European Data Portal

The European Data Portal (EDP) harvests metadata public sector open data portals across European countries. The aim is to improve access and discoverability, and hence facilitate re-use and value creation. The EDP is developed by the European Commission with the support of a consortium led by Capgemini, including INTRASOFT International, Fraunhofer Fokus, con terra, Sogeti, Time.Lex and the University of Southampton.

Following the DCAT and DCAT-AP specifications, EDP considers three main types of artefacts: (i) *catalogues*, which are curated collections of metadata about datasets; (ii) *datasets*, which refer to data published or curated by a single organisation and available for access or download in one or more formats; and (iii) *dataset distributions*, which are made in a specific format (CSV, PDF, RDF, etc).To harvest metadata, the EDP use dataset catalogues and APIs provided by open data publishers. The metadata can be accessed through several interfaces, including SPARQL[11].

EDP has implemented their own Metadata Quality Assessment (MQA) tool, based on a subset of the metrics in [13] and reports on the results of the SHACL validation of the mandatory DCAT-AP properties of the datasets they harvest. As per March 31 2019, all but three of the portals considered had achieved over 90% valid DCAT profiles. We refer the interested reader to the web page of MQA tool for further details.[12] In this study, we will focus on the datasets themselves, including two DCAT recommended properties: `dct:Format` and `dct:Publisher`.

4 Corpus and methodology

4.1 Corpus

Our corpus has two parts. The first part consists of the collection of DCAT-AP catalogues, datasets and distributions harvested by the EDP, available through their SPARQL endpoint. We use this to compute a series of metadata-related metrics. The second part is made of all RDF distributions of all datasets harvested by the EDP. As the EDP stores only links to the distributions, we set up an acquisition process to download the data, including the following steps:

1. Acquire the available metadata of each dataset and its distributions, which are registered with the EDP.
2. Filter datasets having at least one distribution with the label `dct:format` property in the set {*n3, turtle, rdf+xml, ttl, rdf_trig*}. In Section 5.1 we will analyse in detail the different ways publishers used this property.
3. Attempt to download the RDF distributions of datasets extracted in the previous step. We register the Pay-Level-Domain (PLD) of the download URL and store it as the *host* of the dataset. As we will see in Section 5.2, not

[11] https://www.europeandataportal.eu/sparql

[12] https://www.europeandataportal.eu/mqa/

all publishers use the `dct:publisher` property in their metadata, therefore, we had to use the host to get an idea of the publisher.

4. For distributions successfully downloaded in the previous step, parse and validate the RDF using the Raptor RDF library. 2.0.15.[13] Register any parsing or validation errors. Some distributions associated to a dataset represented slices of the same, and we considered them as one distribution in our calculations.

The corpus produced by this methodology has the following known bias factors, that we compare with those of previous studies. (i) Unlike [7, 14], we did not use crawling to construct our corpus. Our approach was similar to [4], which used the LOD cloud DCAT descriptions as a starting point. (ii) We considered only datasets that included in their DCAT description the `dct:format` property. This means we miss some RDF datasets without this metadata. (iii) We did not consider SPARQL endpoints, as it was difficult to determine if they contained several other datasets besides the one linked in the distribution which is indexed by EDP. This means that we might have missed some datasets that do not come with a data dump distribution in RDF. As [4], we do not consider incorrect format tags. (iv) As we approximated publishers using the host's PLD, we might have lost some information about the actual publishers. Sometimes, multiple government agencies pool resources to develop and maintain a joint open data portal to manage economies of scale and encourage knowledge exchange - for instance, a city open data portal might host a dataset published by the local policy department, which is a different organisation than the city council.

We ran the acquisition tool on March 26 2019 and collected 6636 datasets with 8780 RDF distributions. Table 1 provides some descriptive statistics about the corpus. We identified 74 different hosts. The top-10 hosts with most RDF datasets are listed in Table 2. Most host names could be intuitively mapped to a data publisher or local data portal. We noted two PLDs, `dati.opendataground.it` and `nexo.carm.es`, where this is not clear. The former corresponds to the Italian municipality of Albano Laziale, and the latter to the Spanish region of Murcia. Three hosts were from Italy, six from Spain and one from the UK. We also noticed a fewer amount of contributors from France, Norway, Netherlands, Czech Republic, Austria and Finland, and none from catalogs of other EU countries.

Table 1. Descriptive statistics of our dataset corpus

Total datasets	6636	Total distributions	8780
Successful distribution download	8016	Failed distribution downloads	764
Successful distribution validation	6990	Failed distribution validation	1026
Datasets with at least one valid distribution	5856	Triples inspected	137,208,657

[13] http://librdf.org/raptor/

Table 2. Top-10 host domains by number of datasets

Domain	**# (%) of datasets**
www.dati.lombardia.it	2836 (48.4%)
opendata.aragon.es	1252 (21.4%)
dati.opendataground.it (Comune AlbanoLaziale)	1011 (17.3%)
datos.gijon.es	357 (6.1%)
opendata.caceres.es	259 (4.4%)
www.dati.friuliveneziagiulia.it	172 (2.9%)
datos.santander.es	172 (2.9%)
nexo.carm.es (Region Murcia)	126 (2.2%)
opendata.camden.gov.uk	100 (1.7%)
datos.madrid.es	76 (1.3%)
other 64 hosts	275 (4.7%)

4.2 Methodology

We analysed the corpus in terms of uptake, and along three quality dimensions: representational, contextual, and accessibility. We chose the metrics that allowed us to better assess re-usability and interoperability/interlinking, the main keywords of the PSI directive that the EC sought with the adoption of Semantic Web technologies.

Uptake We measured the uptake of Linked Data by comparing the number of datasets in EDP that contained at least one distribution in a relevant format with the number of datasets that included at least one distribution in the following formats {*CSV, TSV, PDF, TXT, XML, XLSX, XLS, ODS, JSON*}, that is, other structured formats, plus PDF and TXT. We chose to ignore files made available as: (i) *ZIP*, as they are often provided as a convenience to download all different distributions in one go; (ii) image formats ({*PNG, JPG*}), as they are mostly used to visualise map data, and cannot be represented in RDF; (iii) APIs, as we did not consider SPARQL endpoints (their natural Linked Data counterparts) in our corpus; and (iv)*HTML*, as in most cases they link to external visualisations or dataset descriptions; and any other format tag. Our intention with this metric is to understand how many datasets are available in RDF with respect to other formats, providing a first measure of interoperability of the dataset landscape.

During our analysis, we noticed that publishers use a range of types as `dct:format` and `dcat:mediaType` values, which are currently not covered by the MQA tool implemented by the EDP. DCAT-AP guidelines recommend the use of the URI file type register operated by the Metadata Registry of the Publications Office of the EU (MRPO) to specify formats/media types.[14] We computed the conformance to this recommendation, and report on the different ways data publishers are assigning this value.

[14] http://publications.europa.eu/resource/dataset/file-type

Representational quality This dimension refers to how well the data is represented in terms of common best practices and guidelines. We considered the following aspects:

1. **Usage of well-known vocabularies** Re-using vocabularies is key for increasing interoperability. Vocabularies for different domains are publicly available and can be found using tools such as Linked Open Vocabularies[15]. In our analysis, a vocabulary was considered to be used by a dataset if a term from that vocabulary appeared in the predicate position of a triple, or in the object position of an `rdf:type` triple. We relied on two sources for vocabularies: the list of from [14], and `prefix.cc` website. We report for each vocabulary the number of valid datasets that use it and the percentage with respect to the total. We also compared the relative percentage of vocabularies in our corpus with the one reported in [14] both for their overall corpus, and for their government datasets.
2. **Usage of proprietary and not well-known vocabularies** Sometimes widely used vocabularies do not provide all the terms required to describe a dataset. Data publishers then resort to creating their own vocabularies to match their needs. Following [14] we considered a vocabulary to be proprietary if is used by only one dataset. However, unlike [14], we did not analyse datasets published by the same host as one dataset, which meant that a vocabulary defined by an organisation used in more than one of its own datasets would not be considered proprietary. Therefore, we also computed the set of hosts associated to each proprietary vocabulary. As a starting point, vocabularies that were not on `prefix.cc` were classified as *not well-known*.
3. **Usage of blank nodes** The scope of blank nodes is limited to the document in which they appear, making them undesirable in Linked Data because they are impossible to re-use and interlink. Therefore, using them in datasets intended for public consumption is not advised. We computed the ratio of blank nodes against data-level constants as in [4, 7]. Given a dataset D, the set of blank nodes in D $\beta(D)$, and the set of data-level constants $dlc(D)$, we defined the blank node ratio as $R = \frac{dlc(D) \setminus \beta(D)}{dlc(D)}$. A higher value of R means fewer blank nodes in D.

We chose these three metrics for the following reasons: usage of (not) well-known vocabularies quantify if publishers are using the vocabularies developed by the EC, and if not, what they are using instead. Use of blank nodes is recognized as limitative of interlinking and reuse. [7]

Contextual quality This category refers to how well datasets were fit for the task at hand. In this category we considered *Provision of provenance information* as indicator. Data provenance helps data consumers understand where the data comes from and who produces it. In the context of open government data, this dimension is particularly important, as publishers in this space usually have an

[15] https://lov.linkeddata.es/dataset/lov

official status. We captured this in two ways: (i) Count the number of DCAT profiles of all datasets registered in EDP that have a `dct:publisher` statement. This provided us with a general overview of how all publishers were using this particular type of metadata. (ii) Count the number of dcat profiles corresponding to datasets in our corpus that included a `dct:publisher` statement. This helped us understand how Linked Data publishers were using this type metadata.

We decided to use only `dct:publisher` as this is the property recommended by DCAT-AP, unlike in [4], who also included `dc:creator`. We did not consider other metrics in this category applied in the literature due to their dependence on the particular information need of the user conducting the search [4].

Accessibility This dimension assesses the relative ease with which both machines and humans can re-use Linked Data resources. Within this space we computed the following metrics:

1. **Dereferenceability of vocabularies** To enable applications to retrieve the definition of vocabulary terms, IRIs should be made dereferenceable. We report this metric for proprietary and not well-known vocabularies. We chose this particular metric as a natural complement to vocabulary usage. If publishers are using their own vocabularies instead of the EC ones, are they at least making them discoverable as well?
2. **Links to external datasets** Links between datasets help data consumers query and explore datasets. From an EDP perspective, being able to combine datasets from different countries is of great value for producing EU-wide aggregations with a single SPARQL query. As [4], we counted an external link for each object's resource IRI in a triple that has a PLD different to the PLD of the host of the dataset. However, contrary to them, we did not check if the IRI was dereferenceable. For each detected external domain, we also computed the number of different hosts that published at least one dataset with an external link to it. We chose this metric as it quantifies the interlinking degree among datasets.

5 Results

5.1 Uptake

971, 160 out of the 1, 426, 804 distributions registered in EDP include the recommended *dct:format* (68.4%). In terms of datasets, 384, 128 out of the 860, 294 contain at least one distribution with declared *dct:format* (44.6%). From these:

1. 476, 377 distributions (44.9%) use the recommended MRPO vocabulary. For the sake of simplicity, we refer to the MRPO namespace as `mrpo`.
2. 127, 015 distributions use *mrpo*, but with a wrong code at the end, *e.g.*, lowercase instead of uppercase, a non-existent format, or combinations of formats in a single IRI (e.g. `mrpo:ZIP+CSV`).

3. 154, 216 distributions used a text literal. Most of them correspond to the codes of common file types.
4. 105, 730 distributions used other IRIs. 96% of them came from one national open data portal that defined for each dataset an instance of the `dcterms:IMT` class with and an `rdfs:label` of the actual format. From a pure DCAT validation perspective, this is correct, as each IRI has the right type. However, this creates unnecessary entities and complicates the querying of different formats, as any aggregation then needs to be done on the text labels.
5. 108, 082 distributions had a blank node, described by an `rdfs:label`. We noticed that Geoportals (portals that hold geographic information) were the most prevalent contributors of this metadata. These portals are aligned to the INSPIRE[16] metadata, specifically designed for geospatial data. In order to integrate metadata about geospatial data with the other types of data, the EC developed the GEODCAT-AP extension, and efforts were undertaken to map INSPIRE to it. According to the documentation, Geospatial data should use the filetypes from the MRPO register as format (or dcat:mediaType), or, in case of absence, use the type register of the INSPIRE project. We suspect that there is an issue on how geoportals export their INSPIRE metadata to GeoDCAT.

Regarding the optional `dcat:mediaType` property, 284, 978 distributions (19.9%), with 114, 990 datasets (13.3%) having at least one distribution with it. 98% of the distributions including *dcat:mediaType* also included *dct:format*. This is good, as DCAT-AP defines the former as a sub-property of the latter. However, similar to *dct:format*, we found that publishers have different ways of setting this value. Some of them use the full URL of the IANA mediaType, *e.g.*,:
http://www.iana.org/assignments/media-types/text/csv
while others chose to use either the registry/name tag ('text/csv'), and a third group used the name ('csv'). According to the DCAT recommendation examples, the registry/name option is the correct one.

[4] measured this in their corpus using the *void:feature* property and found only 9 datasets including it. They recommended extending the metric to include DCAT properties, which is what we did here. However, we hypothesize that due to the existence of the DCAT-AP guidelines, PSI community is more prone to include this property than others.

Given the multiple ways that formats are declared, we decided to count datasets with at least one distribution on format F, with F the case insensitive value text label of each case identified above, i.e., the value of *skos:PrefLabel* for case (1) and (2), the literals for case (3), and the value of *rdfs:label* for cases (4) and (5). Figure 1 shows the comparison per each format. For this sample, RDF is still a minoritary format. Tabular formats (both open and closed) are dominant, in particular CSV (over 100k datasets). RDF is approximately 5 to 6 times less common that non-tabular structured formats like XML and JSON.

[16] https://inspire.ec.europa.eu/metadata/6541

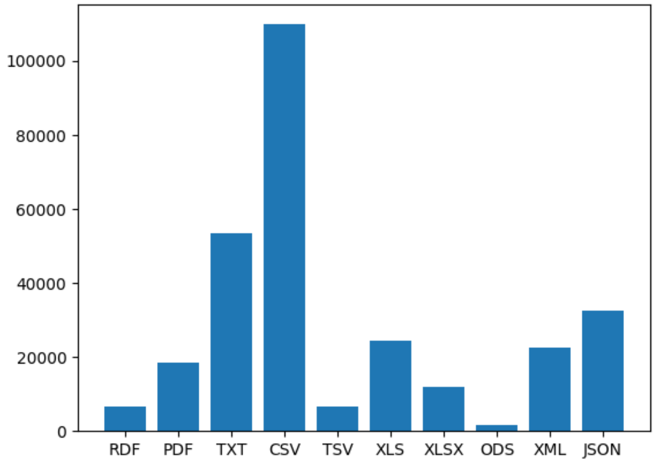

Fig. 1. Number of datasets with at least one distribution on each format

5.2 Basic provenance information

422,058 datasets out of 838,743 have a *dct:publisher* (50.3%). From the 5856 datasets with at least one resource in valid RDF format, we found that 2507 have a *dct:publisher* (42.8%). The latter result was a bit surprising for us, as we expected that organisations that are aware enough of Linked Data technologies to publish their data in RDF, would have their DCAT descriptions as complete as possible, even if *dct:publisher* is considered recommended instead of mandatory by the DCAT-AP specification. However, compared to the corpus in [4], where only 16.27% of datasets included this information, we can say that the PSI community is more committed to add it than others. Given the incompleteness of metadata on publishers, we chose use *hosts* as an estimation of publisher in the rest of our calculations.

5.3 Usage of well-known vocabularies

Table 3 compares the percentage of well-known vocabularies in our corpus with the one reported in [14], both against the overall corpus and against datasets they categorised as Government. Overall, our corpus shows lower re-use percentage for all vocabularies, except for wgs84, when compared to the government slice. The sharpest differences are noted in the use of `dcterms` (-57% wrt (b)), `foaf`(-20% wrt (b)), `dcat` (-28% wrt (b)), `void` (almost nonexistent in our corpus), `cube/qb` (-53% wrt (b)), and `rss` (almost nonexistent in our corpus). The low usage of `dcat` and `void` in our corpus is expected, as by design there are few datasets that describe other datasets (EDP expects one DCAT-AP catalog per portal). The

Table 3. Comparison of quotas of well-known vocabularies detected in [14] between (a) Full corpus of [14] (b) Government subset of corpus of [14] and (c) our corpus. We also add number of hosts and predicates detected in our corpus.

Prefix	**(a)**	**(b)**	**(c)**	**# hosts**	**# Predicates**
rdf	98.22%	98.9%	83.54%	63	10
rdfs	72.58%	95.62%	87.18%	47	9
foaf	**69.13%**	**27.32%**	**7.53%**	43	20
dcterms	**56.01%**	**63.93%**	**6.10%**	47	45
owl	36.49%	23.49%	18.80%	18	5
wgs84	25.05%	7.10%	7.36%	19	4
sioc	17.65%	0.00%	0.00%	0	0
admin	15.48%	0.00%	0.00%	0	0
skos	**14.11%**	**20.76%**	**1.06%**	24	24
void	**13.51%**	**39.34%**	**0.91%**	13	13
bio	12.32%	1.09%	0.00%	0	0
cube/qb	**11.24%**	**61.74%**	**8.69%**	11	16
rss	**9.76%**	**54.64%**	**0.36%**	1	2
odc	8.48%	0.00%	0.00%	0	0
w3con	7.60%	0.00%	0.00%	0	0
doap	6.41%	2.73%	0.00%	0	0
bibo	6.11%	0.00%	0.10%	5	9
dcat	5.82%	28.41%	0.94%	12	12

lesser usage use of `foaf` could be explained by our corpus not including many datasets that talk about people. The same argument could be made for `cube/qb`, as these vocabularies are almost exclusively used for statistical datasets. For `rss`, we believe it has stop being used as it was in 2014, as it is not dereferenceable anymore.

Table 3 also shows the number of different hosts that host data using that vocabulary, and the number of predicates from the vocabulary that are used across all datasets. Interestingly, although the number of datasets is low for `foaf`, `dcterms`, and `skos`, more than 30% of hosts use them. Furthermore, these vocabularies have the highest number of different predicates used.

Table 4 shows the top-10 vocabularies in terms of number of datasets, that exist in prefix.cc and were not already listed in Table 3. `xsd` and `dbpo` are the most prevalent in terms of datasets, while `vcard` and `dc11` the most popular in terms of number of hosts using them. We also found that `xhv` and `opensearch` are used in combination by a single publisher in the same relatively large number of small datasets to describe the results of an informative web page.

5.4 Usage and dereferenceability of other vocabularies

Table 5 shows the most used not well-known vocabularies. Most of them are used by only one host, conforming to the definition of proprietary in [14]. However, only `Aragopedia`, `ontouniversidad` and `server1.avantic.net` were de-

Table 4. Top-10 vocabularies in `prefix.cc` not listed on Table 3

Prefix	**% datasets**	**# Hosts**	**# Preds.**
`xsd`	17.3%	4	1
`dbpo`	13.25%	5	10
`apivc`	13.1%	2	4
`opensearch`	13.1%	1	2
`xhv`	13.1%	1	1
`sprx`	8.3%	1	5
`sdmx`	8.2%	1	3
`vcard`	4.4%	16	39
`dc11`	4.0%	18	31
`geonames`	2.2%	7	8

Table 5. Top-10 not well-known vocabularies by dataset percentage

Vocabulary	**% Datasets**	**# Hosts**	**# Preds**	**Deref-able?**
socrata.com/rdf/terms	52.6%	4	1	No
opendata.aragon.es/def/Aragopedia	13.0%	1	52	No
w3.org/2000/10/swap/pim/usps#	2.7%	4	4	Yes
data.press.net/ontology/stuff/	2.1%	2	5	Yes
opendata.caceres.es/def/ontomunicipio	1.7%	2	139	HTML
purl.org/ctic/infraestructuras/	1.1%	1	5	No
opendata.unex.es/def/ontouniversidad	1.0%	1	63	HTML
dublincore.org/documents/dcmi-box/	0.7%	1	4	No
open.vocab.org/terms	0.6%	1	3	HTML
server1.avantic.net/opendata/vocab/raw/	0.5%	1	206	No

veloped by data publishers (Spanish regions of Aragón, Cáceres, and Cádiz respectively). We highlight the popularity of `socrata.com/rdf/terms` both in number of datasets and different hosts. However, this is not a vocabulary per se: it is comprised of only one property, `socrata:rowID`, that is defined by default by the Socrata open data management tool in its CSV2RDF conversion utility. We can also infer from this that the original format of these datasets is CSV. Interestingly, the 3 predicates used from `open.vocab.org` are `csvHeader`, `csvRow` and `csvCol`, consistent with an attempt to export CSV to RDF. Finally, we note that the `dcmi-box` namespace is incorrect. We believe the publisher meant to use the *dcterms:Box* property.

In terms of deferenceability, only `and` `uspe` returned an rdf+xml description, while `ontomunicipio`, `ontouniversidad` and `openvocab` returned HTML documentation. We are aware that both ontomunicipio and ontouniversidad have RDF versions, so the problem seems to be one of server configuration to return the right representation.

We found more than 3000 proprietary vocabularies, more than 95% of them non-dereferenceable at all. This surprisingly high number is mainly due to what it appears to be an incorrect use of the Socrata's RDF export from CSV utility[17], used by three of the top-10 contributors to our corpus (dati.lombardia.it, datifriulivenziagiulia.it, and opendata.camden.gov.uk). The utility sets a number of namespaces by default, including an auto-generated namespace based on the id of the resource, *e.g.*:

`http://data.cityofchicago.org/resource/xzkq-xp2w/`.

to which CSV headers are appended to create predicates. The default turns out to be quite unhelpful, as a different non-dereferenceable predicate is created for each column of each dataset, yielding an even less interoperable collection than the original set of CSVs. We found that more than 90% of the detected proprietary vocabularies correspond to this pattern. We also found that the Comune AlbanoLaziale portal (based on OpenDataGround[18]) has a similar functionality, that is also configured in a way that generates different predicates per each column header in a per-resource namespace.

5.5 Blank nodes usage

Figure 2 shows the distribution of the blank node ratios of datasets in our corpus. The median is very close to 1, meaning that the majority of datasets have none or almost none blank nodes. However, there is a sizable cluster of 485 outliers with $R \leq 0.1$, that is, 485 datasets with more than 90% of blank nodes.

We took a closer look at those extreme outliers. We found that they were all published by the Aragón region (opendata.aragon.es), as part of the first version of their project `Aragopedia`[19]. Datasets correspond to statistical observations of each of the 485 municipalities of the region. We contacted them about the issue, and they acknowledged that they were aware that the export was indeed faulty, and they were currently working on a fix. It was pointed out to us that both the XML v1 and RDF v2 distribution of these datasets were correct.

5.6 Links to external providers

Table 6 shows the top-10 domains with more datasets linking to them, and the number of different hosts that use them. We also add to the table the `publications.europa.eu` domain to measure the usage in our sample of the controlled vocabularies defined by the EC. *w3.org* and *purl.org* are the most linked to by the most publishers. We also highlight the linking to DBpedia by close to 20% of the publishers, predominantly through the use of common vocabularies/predicates. There is very little linkage to `geonames.org`, which can be considered a bit surprising, as many of the datasets in our corpus are published by regions and municipalities, where we expect data about geographical places or with a spatial dimension. In this study we did not analyse the use of the `dct:spatial` property of DCAT-AP that could be used instead of including spatial statements in the dataset.

[17] https://dev.socrata.com/docs/formats/rdf-xml.html

[18] http://www.evodevo.it/open-data-ground/

[19] https://opendata.aragon.es/aragopedia/

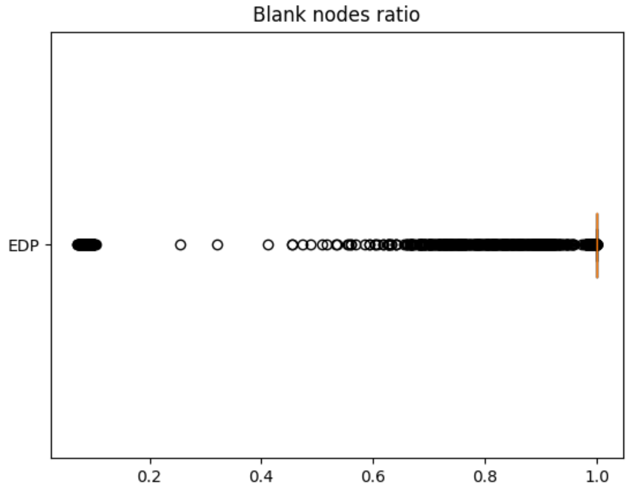

Fig. 2. Distribution of blank node ratio values for datasets in our corputs

Concerning the use of government defined vocabularies, we highlight the presence of *reference.data.gov.uk* in many providers, suggesting that Linked Data publishers do use the definitions in the site: a vocabulary for time intervals, and for defining government offices. However, there is very little linkage both at dataset and publisher level with the *publications.europa.eu* domain, suggesting publishers are not using the controlled vocabularies provided by the EC.

6 Lessons learned and implications

In this paper we assessed the uptake of Linked Data principles on a large sample of open government datasets, made available via the European Data Portal. We measured the popularity of RDF against other publishing formats, and analysed the quality of the RDF datasets according to representational, contextual, and accessibility metrics. We list below main themes that emerged from the findings together with activities planned by EDP to address them to increase the interoperability of open government data:

1. **Recommended DCAT-AP properties are needed to facilitate automated quality analysis.** In our study, the limited use of `dct:format` meant that our sample may have missed some datasets. In the case of `dct:publisher`, we had to consider dataset hosts to identify publishers. Furthermore, the fact that publishers had different preferences for `dct:format` values made querying the data more difficult. The more publishers follow recommended practices, such as using `mrpo`, the easier it is to monitor uptake, identify challenges and propose solutions. In our corpus, we had first to

Table 6. Top-10 domains with more datasets in our corpus linking to them, and number of different hosts where they are published. We add the no top-10 domain publications.europa.eu.

Domain	**# Datasets**	**# (%) hosts**
w3.org	2467	55 (74%)
es.dbpedia	769	4 (5.4%)
purl.org	577	35 (47.3%)
reference.data.gov.uk	504	12 (16.2%)
data.press.net	123	2 (2.7%)
murciaturistica.es	122	1 (1.35%)
geonames.org	119	4 (5.4%)
www.gijon.es	117	1 (1.35%)
schema.org	73	7 (9.5%)
dbpedia.org	43	11 (14.8%)
publications.europa.eu	4	4 (5.4%)

investigate how the property was used (using OPTIONAL clauses, manual inspection etc.) to be able to derive the right query. This makes this sort of analysis more costly than it has to be, which in turn might mean less effective efforts to provide relevant standard updates and guidance.

Follow-up actions EDP will apply SHACL validation to recommended properties and encourage data publishers to follow the recommendations. For `dct:format`, EDP plans to perform the alignment and completion, and share the results with the publishers.

2. **RDF is a minority compared to other structured formats.** Our results suggest that RDF is very seldom the primary format of choice for open government datasets. Most datasets are in tabular format and then transformed to RDF.

   Follow-up actions EDP needs to do more to engage publishers in reviewing the W3C recommendation for generating RDF from tabular data.[20] and kickstart discussions to add the recommendation to DCAT-AP.

3. **Vocabulary re-use is limited** Our results suggest that publishers are having issues finding, using and/or aligning to vocabularies: from the different ways of assigning values to `dct:format`, through the default parameters of Socrata's `csvtordf` conversion, to the low usage of the vocabularies defined by the EC.

   Follow-up actions Considering that lifting from tabular formats appears to be the best way to move forward, EDP will study the feasibility of applying recent research methods in this area [8, 1] to find alignments to tabular headers. This could be done by intermediaries such as the EDP, by portals or the publishers themselves. Centralising the efforts at portal (or meta-portal level, like in EDP) creates economies of scale and guarantees more homogeneous results. Asking the publishers distributes the effort more widely, but

[20] https://www.w3.org/TR/csv2rdf/

additional care needs to be taken in providing clear guidelines for the choice and use of vocabularies, which will always leave some room for interpretation. In addition, portals and meta-portals need to consider the costs of coordinating individual lifting activities.

7 Conclusion and Future Work

In this paper, we conducted a quality assessment of the adoption and quality of Linked Data in the context of the European Data Portal, a portal that indexes European Open Government data. In this context Linked Data is used as a means to improve the re-usability and the interoperability of data assets within the European Union. We found that RDF is still a minority format. Most publishers that provided RDF versions of their datasets do so by taking advantage of capabilities of their portal software to convert CSV or XML datasets into RDF. However, they often do it without providing links to other datasets, or using well-known vocabularies. This suggests a gap between the numerous academic approaches to link CSV files to ontologies, and the tools used by open government data publishers. Besides the *technology readiness* gap, we also believe there is an organisational gap: on the one hand, data publishers may lack the contextual information of what other entities to link to; on the other, portals that only index metadata would need to download and process all datasets. Even if they can produce linksets or RDF versions of the datasets, there is the question how to manage their storage and update.

As future work, in addition to the recommendations outlined in section 6, we would like to categorise the profile of data portal users to apply contextual metrics based on their particular information needs. Our quantitative results could drive the design and execution of a qualitative assessment of the discoverability and fitness for use of datasets in the portal, in the spirit of recent studies on Human Data Interaction in data portals [9]. Finally, we would like to explore the applicability of recent data portal models that integrate social tools common in collaborative software development infrastructures ([12]), to include dataset consumers in the loop with a view to improving dataset quality.

References

1. Alobaid, A., Corcho, O.: Fuzzy Semantic Labeling of Semi-structured Numerical Datasets. In: Faron Zucker, C., Ghidini, C., Napoli, A., Toussaint, Y. (eds.) Knowledge Engineering and Knowledge Management. pp. 19–33. Lecture Notes in Computer Science, Springer International Publishing (2018)
2. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Government Information Quarterly **32**(4), 399–418 (Oct 2015). https://doi.org/10.1016/j.giq.2015.07.006
3. Bischof, S., Martin, C., Polleres, A., Schneider, P.: Collecting, Integrating, Enriching and Republishing Open City Data as Linked Data. In: Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier,

M., Heflin, J., Thirunarayan, K., Staab, S. (eds.) The Semantic Web - ISWC 2015. pp. 57–75. Lecture Notes in Computer Science, Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-25010-6_4

4. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the quality of the LOD cloud: An empirical investigation. Semantic Web **9**(6), 859–901 (Jan 2018). https://doi.org/10.3233/SW-180306
5. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology **1**(1), 1–136 (Feb 2011). https://doi.org/10.2200/S00334ED1V01Y201102WBE001
6. Hendler, J., Holm, J., Musialek, C., Thomas, G.: US Government Linked Open Data: Semantic.data.gov. IEEE Intelligent Systems **27**(3) (May 2012). https://doi.org/10.1109/MIS.2012.27
7. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of Linked Data conformance. Journal of Web Semantics **14**, 14–44 (Jul 2012). https://doi.org/10.1016/j.websem.2012.02.001
8. Kacprzak, E., Giménez-García, J.M., Piscopo, A., Koesten, L., Ibáñez, L.D., Tennison, J., Simperl, E.: Making Sense of Numerical Data - Semantic Labelling of Web Tables. In: Faron Zucker, C., Ghidini, C., Napoli, A., Toussaint, Y. (eds.) Knowledge Engineering and Knowledge Management. pp. 163–178. Lecture Notes in Computer Science, Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-03667-6_11
9. Koesten, L.M., Kacprzak, E., Tennison, J.F.A., Simperl, E.: The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 1277–1289. CHI '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3025838
10. Lopez, V., Kotoulas, S., Sbodio, M.L., Stephenson, M., Gkoulalas-Divanis, A., Aonghusa, P.M.: QuerioCity: A Linked Data Platform for Urban Information Management. In: The Semantic Web – ISWC 2012. pp. 148–163 (2012). https://doi.org/10.1007/978-3-642-35173-0_10
11. Maali, F., Cyganiak, R., Peristeras, V.: A Publishing Pipeline for Linked Government Data. In: The Semantic Web: Research and Applications. pp. 778–792. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2012). https://doi.org/10.1007/978-3-642-30284-8_59
12. Neumaier, S., Thurnay, L., Lampoltshammer, T.J., Knap, T.: Search, Filter, Fork, and Link Open Data: The ADEQUATe Platform: Data- and Community-driven Quality Improvements. In: Companion Proceedings of the The Web Conference 2018. pp. 1523–1526. WWW '18 (2018). https://doi.org/10.1145/3184558.3191602
13. Neumaier, S., Umbrich, J., Polleres, A.: Automated Quality Assessment of Metadata Across Open Data Portals. J. Data and Information Quality **8**(1), 2:1–2:29 (Oct 2016). https://doi.org/10.1145/2964909
14. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data Best Practices in Different Topical Domains. In: The Semantic Web – ISWC 2014. pp. 245–260 (2014). https://doi.org/10.1007/978-3-319-11964-9_16
15. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., schraefel, m.c.: Linked open government data: lessons from Data.gov.uk. IEEE Intelligent Systems **27**, 16–24 (May 2012). https://doi.org/10.1109/MIS.2012.23
16. Villazón-Terrazas, B., Vilches-Blázquez, L.M., Corcho, O., Gómez-Pérez, A.: Methodological Guidelines for Publishing Government Linked Data. In: Wood, D. (ed.) Linking Government Data, pp. 27–49. Springer New York, New York, NY (2011). https://doi.org/10.1007/978-1-4614-1767-5_2