Big Data Quality:
a Roadmap for Open Data

Paolo Ciancarini*   Francesco Poggi*   Daniel Russo*

*Dept. of Computer Science & Engineering
University of Bologna-Alma Mater
{paolo.ciancarini,francesco.poggi5,daniel.russo}@unibo.it

Abstract

Open Data (OD) is one of the most discussed issue of Big Data which raised the joint interest of public institutions, citizens and private companies since 2009. However, the massive amount of freely available data has not yet brought the expected effects: as of today, there is no application that has fully exploited the potential provided by large and distributed information sources in a non-trivial way, nor any service has substantially changed for the better the lives of people. The era of a new generation applications based on OD is far to come. In this context, we observe that OD quality is one of the major threats to achieving the goals of the OD movement. The starting point of this case study is the quality of the OD released by the five Constitutional offices of Italy. Our exploratory case study aims to assess the quality of such releases and the real implementations of OD. The outcome suggests the need of a drastic improvement in OD quality. Finally we highlight some key quality principles for OD, and propose a roadmap for further research.

Keywords

Open Data Quality; Information Modeling; E-Government; Big Data Knowledge Extraction

1. Introduction

The Open Government Initiative is one of the keystone of the Obama administration. In fact, in his presidential inaugural address, on January 21st, 2009 he clearly stated the importance of openness of Open Data. We know that to establish effective big and open data projects we need, in particular, various big data banks and resources with quality data collection, validation, evaluation and certification methods and standards [16]. The OD quality is indeed an issue of Big Data, since we have directly related research such as the exploration of linked and big data guided by ontologies [2]. There is a common agreement about the usefulness of the Open-Linked Data Paradigm as a promising technology for publishing, sharing, and connecting data on the Web, to provide new perspectives for data integration and interoperability [3]. In particular, interoperability issues are becoming more relevant, since the European Court of Justice set recently interoperability itself as a key issue of the development of computer science [12], [11]. The quality issue is often taken into consideration but never investigated in its multifaceted aspects with regard to Open-Linked Data. Interestingly, some authors highlight this issue in the subdomain of Big Data sensing, since it is becoming a new concept and next technology trend based on a connected sensor world because of Internet of Things (IoT) [17].

In the community, there is some consensus to consider Open Data as part of the Big Data landscape [5]. In our gap analysis we propose a sort of case study to explore some key issues regarding Big Data. For instance, the variability of Big Data is reduced by Open Data since the implementation of general agreement (i.e. the adaption of RDF [29] as basic model for OD representation) reduces the complexity of data interpretation. Open-Linked Data are based on a bunch of conventions, e.g., the use of ontologies, created using formal languages i.e. OWL [23]. Since OD are accessed, stored, linked, queried, they can be considered as Big Data in a laboratory setting, where certain variables are under control. So, studying in depth OD quality issues is complementary to deal with the remaining variety dimensions in Big Data at large. According to some relevant literature, the difficulties in making practical use of OD are often attributed to poor quality of OD themselves [21] [24] [25] [26]. In this context, Big Data is a magnifier of the low-level quality issues (e.g. erroneous data, missing data, misuse of basic constructs - such as misleading *owl:sameAs* links, faulty syntax, etc.). Therefore there is an urgent need to explore OD quality issue in order to scale and understand better this problem on a Big Data level.

The positive economic impact of Open Data has been quantified in more than $3 trillion in value every year in several domains of the global economy i.e., Education, Transportation, Consumer Products, Electricity, Oil & Gas, Health care, Consumer Finance [30]. The need for more and good quality OD has led to many Open Data initiatives, assessed by the WWW Foundations with its Open Data Barometer [13]. Still, there is a huge gap between the high capacity countries and the capacity constrained ones. The first ones has established OD policies, generally with strong political backing. They have extended a culture of OD out beyond a single government department with OD practices adopted in different government agencies. So they have increasingly adopted it at a local government level.

Than the Open Data Barometer identifies emerging and advancing countries which have rising or established open data programmers, often as dedicated initiatives, and sometimes built into existing policy agendas, as in Figure 1.

Finally, the capacity constrained countries face challenges in establishing sustainable Open Data initiatives as a result of: limited government; civil society or private sector capacity.

Emerging & advancing

	Country	Score	Rank	Change
Top 5	Spain	59.89	13	↑ 4
	Chile	58.7	15	↑ 10
	Czech Republic	58.07	17	↑ 5
	Brazil	52.13	21	↑ 7
	Italy	50.58	22	↓ -2
Bottom 5	Costa Rica	31.26	41	↓ -5
	Tunisia	28.57	45	↑ 5
	China	28.12	46	↑ 15
	Philippines	23.19	53	↓ -6
	Morocco	21.11	55	↓ -15

Cluster avg	**Top dataset**	**Bottom dataset**
	Census data	Land ownership
	Greatest impact	**Least impact**
	Accountability	Inclusion

OpenData Barometer

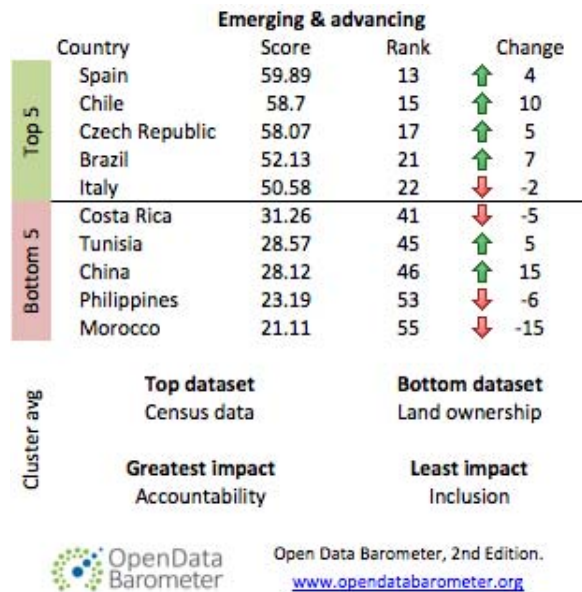Open Data Barometer, 2nd Edition.
www.opendatabarometer.org

Fig. 1. Open Data Barometer 2015: Emerging Advancing Countries

This due to the limits on affordable widespread Internet access, and weaknesses in digital data collection and management.

The OD quality is for all such country and economic sectors a priority for enabling economic growth and citizen participation.

This paper is structured as follows: In section 2 we describe the scenario of our case study. In section 3 we outline the guidelines and standards used for OD implementations by the Italian public administrations. Section 4 is a review about Data Quality Models, with a particular focus on Web portals. In Section 5 we carry out the analysis about OD quality within the five highest offices in Italy. Finally, we draw out conclusions in Section 6 and outline a roadmap.

2. The Italian Case Study

Italy is, according to the Open Data Barometer [13] among the top 5 emerging countries regarding OD. To investigate OD quality we found Italy an interesting case study since it is an average country in the use of OD [13]. It is not a capacity constrained one, where the major problems lays down to a poor internet connectivity and it is not a high capacity one with a high level of expertise routines [32]. Thus, to carry out an exploratory case study on OD quality we found this country interesting and representative, since it face all the major problems related to OD quality.

Open Data represents a subdomain of Big Data. Thus, our approach was to analyze the quality of this representative subdomain, in order to draft our roadmap. Since our analysis addresses OD, it is also meaningful for the more extensive domain of Big Data.

Moreover, OD are spread especially by public administra-

tions to develop application ans services regarding the query of data. This huge potential, estimated by [30], is a great opportunity for the community. Nevertheless, quality regarding OD still represents an unsolved issue. Therefore we tackle this aspect directly in section 5.

Open Data is a clear priority of the Italian Government, and in particular of the Italian Digital Agency (AgID), the governmental office which has as main goal the implementation of the Italian Digital Agenda[1]. Moreover it contributes to the widespread use of information and communication technologies, encouraging innovation and economic growth.

In particular, art. 9 of the Decreto Legge 179/2012 rewrote the art. 52 of the Digital Administration Code (CAD) regarding the open access of public available data. One of the most relevant provision states the *Open Data by default* principle. This pivotal principle states that each act of the public administration is set as open by default, and is considered by AgID itself as one of the most relevant progress in legislation [33]. So, the leveraging effect of this disposition has a huge impact on the future volume of OD.

Italy has still a jeopardized framework regarding OD. More in detail, any office use to have its own internal regulation about how to deliver OD. There is an urgent need to systematize and homogenize on national level the process and standard of OD release. According to AgID, all offices have to comply with its guidelines, according to W3C international standards, since Open Data belongs to the community [33].

3. AgID's standard for Open Data

OD are innovation drivers, thus need a set of standards and process to be easily implemented within the whole Italian public administration [33]. In detail, for a *ex ante* analysis:

- data has to be on an open format to be available. This means with a clear ontology and neutral with respect to the used technology;
- the license should permit the most open use, also for commercial use and in a disaggregated format;
- full accessibility with any automated technology with all metadata should be guaranteed;
- free of use through public or private technological platforms should be granted.

Moreover, AgID states [33] that Open Data, in order to be considered "Open" (for a *ex post* analysis), has to comply with:

- availability: all OD have not to be classified or privacy sensitive. If data do not harm these conditions, they are public and may be available as Open Data;
- accessibility: they have to comply with the EU Directive 2003/98/CE, recently modified by the Directive 2013/98/CE. So, data should be automatically processable with metadata;
- free: all data should be given for free or at the marginal cost.

---

1. http://www.agid.gov.it/agenda-digitale

TABLE 1. AgID's 5 star model

	1 Stars	2 Stars	3 Stars	4 Stars	5 Star
Service Level	No service	Inefficient services	Inefficient services	Efficient services	Efficient services - Data Mash-Up
Accessibility	Human only	Human and semi automatic	Human and semi automatic	Human and automatic	Human and automatic
Information	Document	Raw data	Raw data	Semantic enriched data	Semantic and linked enriched data

TABLE 2. Metadata Framework

	Absence of metadata	External metadata related on a dataset	Internal metadata related on a dataset	Internal metadata related on a data
Level of detail	None	Dataset	Dataset	Data
Relation data-metadata	Absent	Loose	Strong	Strong

The Berners-Lee 5 star model[2] is given as standard for data and metadata, since it is the most accepted framework [33]. According to AgID, this model is the reference model to classify OD in Italian public administrations

Since the 5 star mode led to some confusion in the past, AgID widened its means, as shown in Table 1. In particular, data of the public administrations have:

1) 1 star if published in any format with an open license (typically PDF). No service level is provided and the document is not processable automatically;
2) 2 stars if published in any automatic processable format with a proprietary license (e.g. Excel). The service level is not efficient because only raw data are available;
3) 3 stars if published in any automatic processable format with an open license (e.g. csv, geoJSON);
4) 4 stars if published on a 3 star level but using W3C RDF or SPARQL standards. Those data have an efficient service level, they are fully automatically processable and data are enriched semantically;
5) 5 stars if published on a 4 star level but containing also links to external OD. There is Data Mash-Up and data are linked with external links, to identify better the data context.

Clearly, to generate a new level of information through information inference, all OD themself should be Linked Data. AgID stresses this element very clearly and propose as best case the one of the Italian Parliament [33]. Since metadata are fundamental to generate Linked Data, AgID proposes also its quality model, represented in Table 2.

All those classifications should lead and guide public administrations in Italy to release their OD. For sure they are widely recognized criteria but apparently hardly implemented. Our analysis will show the gap between common recommendations and standards and the real OD implementation of the five highest Italian public offices.

2. More information are available online at 5stardata.info.

4. Data Quality Models

Data quality models are widely known in software engineering [34]. Often, data quality is defined as the ability of a data collection to meet user requirements [8]. Literature on data quality began in the context of information systems [35]. It has been widened to various fields in computer science, such as cooperative systems [14], data warehouses [6] or e-commerce [1]. Research about data quality on the Web, due to the differences from the traditional information systems [34] started with an independent approach [19]. In fact, in a systematic literature review, there was no work on Web data quality [9]. This is quite curious, since data quality is one of the most relevant factors in the quality of a Web portal [37].

The organization and representation of data was a priority well before Open Data. A useful survey shows how, within the four dimension of Intrinsic, Contextual, Representational, and Accessibile information, quality- literature interprets the issue of data quality [28].

Recently, also ISO standards introduced data quality assessments within 2501n family which is focused on software product quality models. So, the 25012 standard [ISO/IEC-FDIS-25012] is focused on data. This standard defines fifteen dimensions, classified in two main dimensions: inherent and system dependent. Inherent is whenever data have the intrinsic characteristic to satisfy needs when data is used under specified conditions (e.g., data domain, metadata, relationships). On the other hand, system dependent refers to the degree to which data quality is attained and preserved within a computer system when data is used under specified conditions. From this point of view, data quality depends on the technological domain in which data are used (e.g. obtaining the required precision in hardware devices, assuring recoverability in backup software, etc.). Table 3 summarizes all those dimensions within the ISO/IEC 25012 framework.

ISO/IEC 25012 should, so:

1) define and evaluate requirements in the production, acquisitions and integrations of data;
2) identify the criteria for quality assurance of the data;
3) evaluating data compliance with national laws and / or existing requirements.

However, there is still no consensus about data quality released on Web portals. There are some attempts in literature to systematize and propose quality models for the Web, but they are still quite isolated [10] [31].

Nevertheless, in order to run our case study, we consider some of the proposed principles and investigate about the implementation of the five highest offices in Italy.

TABLE 3. ISO/IEC-FDIS-25012

CHARACTERISTICS	DATA QUALITY	
	INHERENT	SYSTEM DEPENDENT
Accuracy	X	
Completeness	X	
Consistency	X	
Credibility	X	
Currentness	X	
Accessibility	X	X
Compliance	X	X
Confidentiality	X	X
Efficency	X	X
Precision	X	X
Traceability	X	X
Understandability	X	X
Availability		X
Portability		X
Recoverability		X

5. Analysis

As discussed in the previous section, evaluating the quality of data in the web context, and in particular of OD, is not a trivial task. Different models, standards and approaches have been proposed, but no one provides a complete and widely accepted framework for estimating the OD quality, and driving their lifecycle so as to ensure data producers that their products and services consistently meet consumers requirements. This problem is one of the major threats and key point to achieving the goals of the OD movement.

In this section we present an analysis of the quality of OD released by the five highest offices in Italy. The main information of these datasets is summarized in Table 4. The status of the OD of the Italian highest organs is quite fragmented and heterogeneous. The worst situation concerns the Presidency of the Republic, which has a traditional website containing news about the President activity only in textual formats (e.g. PDF).

The Constitutional Court and the Presidency of the Council of Ministers have a website dedicated to OD. The first contains personal data about the judges of the court and all their judgements, since 1956, in XML format. The latter includes more heterogeneous data concerning the activity of the Council of Ministers, with information spanning from personal data to the acts, contributions, grants, expenses, contracts, etc. of its members. In this case, data are provided in textual (e.g. PDF) or tabular (e.g. XLS) formats. Both datasets are quite small, with a total size ranging from 348 to 679 MB, and have 3 on 5 stars in Berners-Lee deployment scheme for OD.

The last two portals provides access to the OD of the Parliament of the Italian Republic. They contain information (e.g. personal data, election results, group composition, legislative acts, votes, etc.) about the two parliamentary houses (i.e. the Chamber of Deputies and the Senate of the Republic) which

---

3. http://www.quirinale.it/

4. http://www.cortecostituzionale.it/ActionPagina_1177.do

5. http://www.governo.it/AmministrazioneTrasparente/

6. http://dati.camera.it/

7. http://dati.senato.it/

---

compose the Italian parliament. Both of them provide a huge amount of data in RDF [29], modeled using two OWL [23] ontologies[8]. These datasets can be downloaded, or queried online using the SPARQL [22] endpoints provided by the two agencies. In addition, the dataset of the Senate of the Republic is available for download also in other non-semantic formats (i.e. JSON, CSV and XML). Under the Berners-Lee five star scheme, both datasets have 4 on 5 stars, since they use open standards from W3C such as RDF and SPARQL to identify things (the fourth star principle), but they don't contain links to other people's data (fifth star principle).

Another interesting analysis is testing how the datasets in the five OD portals behave with respect to the data quality dimensions defined in the ISO 25012 standard. While basic quality dimensions like *conformance* to accepted standards, *credibility* (i.e. data come from trustworthy sources), *processability* (i.e. data are provided in machine-readable formats), etc. are fully fulfilled, others quality requirements are only partially met. The *timeliness* dimension, for instance, is particularly critical for all the portals under examination. Let us consider the following query, which requests the list of the Presidents of the Italian Republic, ordered by mandate date:

```
PREFIX ocd: <http://dati.camera.it/ocd/>

SELECT DISTINCT ?title ?date
WHERE {
   ?s a ocd:presidenteRepubblica;
        dc:date ?date ;
        ocd:rif_persona/dc:title ?title
} ORDER BY ?date
```

The query has been submitted to the SPARQL endpoint of the Chamber of Deputies on Feb. 2015, getting the results showed in Table 5. The actual President of the Republic, Sergio Mattarella, which has been elected on Feb. 3rd 2015, has not been included in the dataset.

TABLE 5. The list of the Presidents of the Italian Republic, ordered by date. The dataset is outdated of 9 months, at least.

NAME	DATE
ENRICO DE NICOLA	19480101-19480512
LUIGI EINAUDI	19480512-19550511
...	...
CARLO AZEGLIO CIAMPI	19990518-20060515
GIORGIO NAPOLITANO	20060515

In addition to these limitations, there are also problems that concern the quality of the data model (e.g. the schema, the ontology, etc.) used to represent the information contained in structured datasets, and that hinder data full exploitation. Although several techniques [7] and metrics [18] have been proposed to evaluate the quality and expressivity of models

---

8. The ontologies of the Chamber of Deputies and of the Senate of the Republic are available at http://dati.camera.it/ocd/classi.rdf and http://dati.senato.it/application/xmanager/projects/datisenato/file/osr.rdf, respectively

TABLE 4. A summary of the OD portals of the five Italian highest offices

	Size	Formats	SPARQL endpoint	# of triples	# of stars
Presidency of the Republic[3]		PDF, HTML	N	0	1
Constitutional Court[4]	348 MB	XML	N	0	3
Presidency of the Council of Ministers[5]	679 MB	PDF, DOC, XLS, CSV, HTML	N	0	3
Chamber of Deputies[6]	147GB (in RDF/XML)	RDF	Y	125501148	4
Senate[7]	50GB (in RDF/XML)	XML, JSON, CSV, RDF	Y	40292304	4

in the ontology domain, none of them have been considered by the data quality models currently available. Nonetheless, these aspects are crucial in the context of OD, since limitations to data structure and representation poses serious limits to information re-use and inter-operability [4], which are the main objectives of OD.

In the information engineering domain, best practices and conventions for limiting and mitigating these problems are documented in literature [20] [36]. The following list summarizes some of the most evident limitations of the dataset of the Chamber of Deputies:

- *poor compliance with standards/conventions:* class names should be in mixed case, with the first letter of each internal word capitalized. The ontology of the Chamber contains, for example, defines the classes *ocd:documento* (document), *ocd:luogo* (place), etc.;
- *repeated information:* (probably) due to problems in the data conversion process, the dataset contains repeated data. For instance, all the information about the type of human agents are repeated twice;
- *flat structure:* best practices suggest to organize information in a hierarchical structure. The ontology of Chamber of Deputies introduces 71 new concepts and 89 properties. Only 3 concepts are subclasses of other classes, and 7 properties are subproperty of other ones;
- *limitations to data re-use:* best practices in knowledge engineering encourage the reuse of popular and high-quality standards. Au contraire, the Chamber of Deputies ontology includes only the FOAF ontology[9] to represent personal information. No other ontology modules are included. This limits data reuse, or at least imposes additional (usually manual) effort to analyze and understand the data model;
- *missing (re-)use of best practices:* design patterns are recurring solutions to common problems in specific contexts. Ontology Design Patterns (ODPs) [15], in particular, are an emerging technology that favors the reuse of encoded experiences and good practices in ontology design. The presidency of the Chamber of Deputies or the Republic, for instance, are two examples in which the AgentRole or RoleInTime pattens can be used with success. Unfortunately, no one of these patterns have been considered while developing the Chamber of Deputies ontology. This produces the unpleasant and anti-intuitive

effect that, if one users interrogates the SPARQL endpoint asking for the President of the Chamber of Deputies, the results includes some people multiple times (i.e. those elected multiple times), dead people, etc.;
- *errors in the model:* 7 classes contain formal errors in their definition. For example, the class *ocd:adesioneGruppoMisto* defines an incomplete restriction definition. For this reason, standard libraries and tools such as reasoners have problems (e.g. produce exceptions and errors) to manage this dataset.

Due to lack space, we omit here a discussion of the Senate data model. Nonetheless, the illustrated observations are the same for the Senate dataset, since the ontology is very similar, probably because it has been developed by the same team.

6. Discussion and Conclusions

We conducted an exploratory case study about the quality of Open Data in Italy. We took into consideration the five highest offices in Italy. From the analysis we conclude a quite jeopardized ecosystem. While the Senate and the Chamber of Deputies has a good release quality (even with some loopholes), other offices seems to have no valuable OD and release methodology.

Our analysis highlighted that existing data quality models (e.g. ISO/IEC 25012, 5 star model) are not sufficient to measure all those aspects that are relevant in the OD context. In fact, as described in Section 5, even those datasets that have a good quality level within the presented frameworks contain numerous limits and problems, posing serious threats to their effective (re-)use.

Future research should tackle, in particular, the methodological aspect of the release. In fact, even the Senate and the Chamber of Deputies, as forerunners, do not use a recognizable methodology. This, due the fact that there is no common agreement about a development and deliver procedure. These effects will also be magnified with the release of OD by regional governments and municipalities. Confusion about the release process could have tremendous effects, like the release of sensitive data or, in the best case, the failure to release OD, or limits to their re-use.

Moreover, an organic and comprehensive metric framework shoud be developed to support high quality OD release and deployment. In particular, well-established software engineering approaches based on product and process metrics should be used. For what concerns product metrics, existing data quality metrics should be refined, extended and integrated

9. The Friend Of A Friend (FOAF) language is available at http://xmlns.com/foaf/spec/

with model metrics in order to take into consideration the dimensions and characteristics specific to the OD domain. Traditional process metrics [27] should also be considered to guide (i.e. by characterizing, evaluating, predicting, improving, etc.) the definition of effective methodologies and development processes for OD production.

As future work, we plan to focus on the review of OSS available tools to help the OD release process. Moreover, a workflow, based also according to the best cases, will be proposed. After that an analysis about how this tools can be integrated with this workflow will be carried out. Finally we will focus also on (semi-)automatized tool development to support the OD lifecycle. This, for overcoming the lack of specific tools for the workflow. The final result will be a comprehensive proposal about for the Open Data release and methodologies for Italian public administrations.

Acknowledgments

References

[1] M Aboelmeged. A soft system perspective on information quality in electronic commerce. In *Proceeding of the Fifth Conference on Information Quality*, 2000.

[2] Periklis Andritsos and Patrick Keilty. Level-wise exploration of linked and big data guided by controlled vocabularies and folksonomies. *Advances In Classification Research Online*, 24(1):1–3, 2014.

[3] Sören Auer, Theodore Dalamagas, Helen Parkinson, François Bancilhon, Giorgos Flouris, Dimitris Sacharidis, Peter Buneman, Dimitris Kotzinos, Yannis Stavrakas, Vassilis Christophides, et al. Diachronic linked data: towards long-term preservation of structured interrelated information. In *Proceedings of the First International Workshop on Open Data*, pages 31–39. ACM, 2012.

[4] Gioele Barabucci, Angelo Di Iorio, and Francesco Poggi. Integration of legal datasets: from meta-model to implementation. In *15th International Conference on Information Integration and Web-based Applications & Services (IIWAS 2013)*, pages 585–594. ACM, 2013.

[5] Christian Bizer, Peter Boncz, Michael L Brodie, and Orri Erling. The meaningful use of big data: four perspectives–four challenges. *ACM SIGMOD Record*, 40(4):56–60, 2012.

[6] Mokrane Bouzeghoub and Zoubida Kedad. Quality in data warehousing. In *Information and database quality*, pages 163–198. Springer, 2002.

[7] Janez Brank, Marko Grobelnik, and Dunja Mladenic. A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170, 2005.

[8] Cinzia Cappiello, Chiara Francalanci, and Barbara Pernici. Data quality assessment from the user's perspective. In *Proceedings of the 2004 international workshop on Information quality in information systems*, pages 68–73. ACM, 2004.

[9] A Caro, Coral Calero, Ismael Caballero, and Mario Piattini. Data quality in web applications: A state of the art. In *IADIS International Conference WWW/Internet*, volume 2, pages 364–368, 2005.

[10] Angélica Caro, Coral Calero, Ismael Caballero, and Mario Piattini. A proposal for a set of attributes relevant for web portal data quality. *Software Quality Journal*, 16(4):513–542, 2008.

[11] Paolo Ciancarini, Daniel Russo, Alberto Sillitti, and Giancarlo Succi. A guided tour of the legal implications of software cloning. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 2016.

[12] Paolo Ciancarini, Daniel Russo, Alberto Sillitti, and Giancarlo Succi. Reverse engineering: a legal perspective. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. ACM, 2016.

[13] World Wide Web Foundation and Open Data Institute. Open data barometer global report (second edition). http://www.opendatabarometer.org, 2015. Accessed: 2011-02-11.

[14] M Fugini, Massimo Mecella, Pierluigi Plebani, Barbara Pernici, and Monica Scannapieco. Data quality in cooperative web information systems. *Personal Communication. citeseer. ist. psu. edu/fugini02data. html*, 2002.

[15] Aldo Gangemi and Valentina Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.

[16] J. Gao, S. Yu, and L. Cao. Message from the ieee bigdataservice 2015 program chairs. In *IEEE First International Conference on Big Data Computing Service and Applications (BigDataService)*, pages xii–xii, 2015.

[17] Jerry Gao, Lihui Lei, A Guruvayya, MK Vijaykumar, RR Balasubramanian, and V Patil. Big data sensing and services: A tutorial. In *IEEE First International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 79–88. IEEE, 2015.

[18] Juan García, Francisco JoseGarcía-Peñalvo, and Roberto Therón. A survey on ontology metrics. In *Knowledge management, information systems, E-learning, and sustainability research*, pages 22–27. Springer, 2010.

[19] Michael Gertz, M Tamer Özsu, Gunter Saake, and Kai-Uwe Sattler. Report on the dagstuhl seminar. *ACM SIGMOD Record*, 33(1):127–132, 2004.

[20] Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media, 2006.

[21] Harry Halpin, Patrick J Hayes, James P McCusker, Deborah L McGuinness, and Henry S Thompson. When owl: sameas isnt the same: An analysis of identity in linked data. In *The Semantic Web–ISWC 2010*, pages 305–320. Springer, 2010.

[22] Steve Harris, Andy Seaborne, and Eric Prudhommeaux. Sparql 1.1 query language. *W3C Recommendation*, 2013.

[23] Pascal Hitzler, Bijan Krtzsch, Markus amd Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. Owl 2 web ontology language primer. *W3C Recommendation*, 2012.

[24] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In *International Workshop on Linked Data on the Web (LDOW2010) at WWW2010*. CEUR, 2010.

[25] Prateek Jain, Pascal Hitzler, Peter Z Yeh, Kunal Verma, and Amit P Sheth. Linked data is merely more data. In *AAAI Spring Symposium: linked data meets artificial intelligence*, volume 11, pages 82–86, 2010.

[26] Krzysztof Janowicz. Place and location on the web of linked data. http://stko.geog.ucsb.edu/location_linked_data, 2013. Accessed: 2016-02-11.

[27] Stephen H Kan. *Metrics and models in software quality engineering*. Addison-Wesley Longman Publishing Co., Inc., 2002.

[28] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. Aimq: a methodology for information quality assessment. *Information & management*, 40(2):133–146, 2002.

[29] Frank Manola and Eric Miller. Resource description framework (rdf). primer. *W3C Recommendation*, 2004.

[30] J Manyika, M Chui, D Farrell, S Van Kuiken, P Groves, and EA Doshi. *Open data: Unlocking innovation and performance with liquid information*. McKinsey, 2013.

[31] Carmen Moraga, Mª Ángeles Moraga, Angélica Caro, and Coral Calero. Spdqm: Square-aligned portal data quality model. In *Ninth International Conference on Quality Software, QSIC*, pages 24–25, 2009.

[32] Richard R Nelson and Sidney G Winter. *An evolutionary theory of economic change*. Harvard University Press, 2009.

[33] Agenzia Digitale per l'Italia (AgID). Linee guida nazionali per la valorizzazione del patrimonio informativo pubblico. http://www.agid.gov.it/sites/default/files/linee_guida/patrimoniopubblicolg2014_v0.7finale.pdf, 2014. Accessed: 2016-02-11.

[34] Roger S Pressman. *Software engineering: a practitioner's approach*. Palgrave Macmillan, 2005.

[35] Diane M Strong, Yang W Lee, and Richard Y Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[36] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi. *Ontology engineering in a networked world*. Springer Science & Business Media, 2012.

[37] Zhilin Yang, Shaohan Cai, Zheng Zhou, and Nan Zhou. Development and validation of an instrument to measure user perceived service quality of information presenting web portals. *Information & Management*, 42(4):575–589, 2005.