Journal of Computer Science

Original Research Paper

An ADD-Oriented Software Architecture for Structuring Information to Open Government Data

[1,3]Andreiwid Sheffer Correa, [1]P.L.P. Correa, [2]F.S.C. Silva and [3]T. Carvalho

[1]Department of Computer Engineering and Digital Systems, Universidade de São Paulo, São Paulo - SP, Brazil
[2]Institute of Mathematics and Statistics - Universidade de São Paulo, São Paulo - SP, Brazil
[3]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Campinas - SP, Brazil

Article history
Received: 08-12-2017
Revised: 02-02-2018
Accepted: 8-03-2018

Corresponding Author:
Andreiwid Sheffer Correa
Department of Computer Engineering and Digital Systems, Universidade de São Paulo, São Paulo - SP, Brazil
Email: andreiwid@gmail.com

Abstract: The huge number of heterogeneous and not standardized websites where public records are disclosed is an evidence of how unprepared governments are about Open Government Data movement. This scenario is seen all over the world, especially in organizations which open data principles are new e.g. local governments. Thus, not effective law enforcement and the lack of openness result in a waste of resources applied to useless publication of documents, making it hard to achieve the benefits of open data. Looking for a way to help government at this hardworking task, this paper proposes a software architecture for structuring information into open data from distributed repositories placed at local governments websites. This work follows Attribute-Driven Design (ADD) methodology, where the requirement analysis for proposed architecture has been conducted based on availability problems occurred in 561 websites from Brazilian cities. The main contribution of this paper is the proposition of an architecture whose dissemination and instantiation will allow society to focus efforts on more important tasks e.g. review data consistency, making it possible in a near future to automatize the structuring process in a way that data can be used for any purpose.

Keywords: Open Data, Open Government Data, Software Architecture, Structuring Process

Introduction

The emergence of legislation aimed at guaranteeing the entire process of access to public information is gaining ground in modern democracies and introduces a diversity of dedicated websites or data portals (Relly and Sabharwal, 2009). In this context, the Internet has contributed enormously to enabling the use of Information and Communication Technologies (ICTs) to satisfy society for information and to materialize the interest of governments in becoming open and dialoguing with society (Bertot *et al*., 2010; Smith, 2010).

The movement called Open Government Data (OGD) has emerged and established several conceptual and technical principles aimed at guiding the opening of public records using ICTs. Basically, for an initiative to be considered in accordance with the principles of OGD, this must follow certain requirements where data can be freely used, reused and redistributed by anyone, for any purpose (Tauberer, 2014).

The availability of data according to OGD increases the possibility of analysis of public records by allowing them unrestricted access. This means that there is greater participation of society, usually represented by critics as journalists, economists, political scientists and others. The challenge of these professionals would be directed to critical task of data understanding and analysis and not the technical processing or treatment to make it understandable.

After unprecedented initiatives towards OGD introduced by Barack Obama in the US Government in 2009, there was another important step towards the opening of public records with the Open Government Partnership (OGP), in which Brazil was one of eight founding members. The main effect of Brazilian participation in OGP was the development and implementation of its national Access to Information Law (CGU, 2011).

Even considering these important advances, there is still much to do to transform documents into data. One of the main problems is the lack of preparation and knowledge about the OGD principles. This is most found in local governments (Corrêa *et al*., 2014; 2017).

In this sense, there is a proliferation of transparency websites that are indeed repositories of documents like printed reports, usually in PDF and HTML (Corrêa *et al*., 2014; 2017). Thus, unrestricted access to data is compromised given the technical limitations imposed by these formats.

This work aims to define a software architecture to allow the structuring of documents arranged in the existing websites and transparency repositories, with the aim of making them unrestrictedly accessible to specialized critics without the technical barriers that require the processing of formats. In this study we followed the Attribute-Driven Design (ADD) develop by Software Engineering Institute (Bass *et al*., 2012), which is an industry-leading methodology for building software architecture.

After this brief introduction, we emphasize the contribution of this paper and present some works related to this research. Then we present the quality requirements obtained from the analysis based on problems found in 561 Brazilian cities' websites and translate them into architecture requirements following ADD (Bass *et al*., 2012) method. In the ending, we envision the use of structured data in the digital forensics analysis and discuss some perspectives to future works.

Contribution

The main contribution of this research to Computer Science is the definition of a software architecture as a solution model for data availability problems even though data sources are not yet adequate according to the OGD principles. The definition process of a software architecture is described in this paper with the support of the ADD (Bass *et al*., 2012) method.

Currently, we see in government agencies several heterogeneous websites where content is based on the disclosure of documents instead of data. In this context, technical staff responsible for disclosing public information usually do not know OGD and their role is more in line with content publishing. In an ideal situation where there is the existence of data portals that comply with open data, the technical teams know and implement OGD and their role are directed to data science.

Between the current and the ideal situation there are different stages to evolve towards OGD. This evolution is stimulated by full redesign of existing work processes whose demand comes from society itself.

As full redesign of work processes is a time-consuming task and may involve a lot of resources in order to consider OGD since data creation, there is an opportunity to establish an intermediate stage where both public agencies and society can benefit. In this context arises the architecture proposed by this research as a model of propelling solution so that the society perceives the benefits of OGD while agencies' technical teams promote organizational changes in the work processes.

In addition, this research also contributes with a process model for building reference software architecture. This process model was based on methods used by industry which comprise the construction and validation of architectures.

Related Work

Several works on literature define concepts and approaches directly related with this work. The first that we mention is the Maturity Model for Open Data proposed by British Government (Dodds and Newman, 2015). Despite being conceptually defined as a maturity model to be applied in institutional assessments, instead of a software architecture. Authors' model focusses on general requirements that any organization should comply with and on guided processes that should be incorporated. Theses process are similar to prescriptions on how to model an open data infrastructure, despite they are not prescriptive. The British model also highlights that models can be used as reference for organizations create and evaluate theirs actions for open data. However, it is important to observe that British model does not propose a solution for unstructured data sources. It just provides a tool to identify and evaluate open data in the government agencies.

In a work proposed by Kong (2016), the author describes a solution for open data availability that deals with interoperability and which is based on an architecture oriented to services and open standards. The architecture can be considered like the one proposed here, but it does not consider legacy and unstructured data sources that exist within agencies.

Pires (2015) refers to technical recommendations of World Bank to report models of open data infrastructure implementation to keep databases accessible and updated. Even not been a fully software architecture, it proposed technical solutions to consider the coexistence of different database implementations, including the ones which models are characterized by unstructured data sources. Recommendations include three solution models that consider the number of pre-existing databases. However, we highlight that World Bank recommendations are succinct and did not propose any software or technological solution, focusing just on simple flowcharts with elements recommended to promote data availability. Furthermore, World Bank proposals do not consider technical limitation of agencies.

Machado and Oliveira (2011) proposed a standard architecture for agency website in a way to establish a semantic agreement among heterogeneous data sources, allowing data fusion in connected open data scenario (Linked data). This architecture is defined using layers and specifically one of these layers is designed for dealing with semi-structured and not structured data. Another important feature of Machado and Oliveira's work (2011) is the semantic mapping based on ontologies used to map

data into tipple RDF (Brickley and Guha, 2014) and persist it in a semantic database contained in the same layer context. This research approach is an ideal scenario for open data, but there are no directions on how to achieve it considering the legacy and unstructured data sources.

Quality Requirements

Architectures enable building systems to satisfy Architecturally Significant Requirements (ASR), which should deeply affect target system itself and satisfy business goals. The ASRs can be raised using different approaches. To interpret the business model, conducting interviews or even performing workshops with stakeholders are just a few examples of this process (Bass *et al*., 2012; Kazman and Cervantes, 2016).

To identify ASRs, this paper relied on previous published results by Corrêa *et al*. (2017) which collected and evaluated data availability problems occurring in 561 Brazilian cities' websites. Fig. 1 shows main data formats present in evaluated websites and Fig. 2 depicts technical requirements for data availability.

Table 1 below expresses the ASRs identified using the Utility Tree tabular format proposed by (Bass *et al*., 2012; Kazman and Cervantes, 2016). After describing the ASRs, they are evaluated based on two criteria that represent the dimensions "value for the business" and "architectural impact". The evaluation is expressed by the values (High) H, (Medium) M and (Low), L.

For the first dimension, the value for the business, "H" indicates that the architecture must have the requirement; "M" indicates that the requirement is important but does not lead to project failures; "L" indicates that it is a requirement with little value for the business, but it may not be addressed by the architecture if a great deal of effort is required to satisfy it.

For the second dimension, the architectural impact, "H" indicates that its satisfaction profoundly affects the architecture; "M" affects in some way; "L" indicates little effect.

Architecture Requirements

Based on the identified quality requirements, a four-layered reference architecture was developed from the initial proposal of a collaborative-oriented middleware (Correa *et al*., 2014; Corrêa *et al*., 2015) located in the context of the architecture which allows the structuring of information found in distributed repositories and make them available in OGD repository. It is important to emphasize that for the scope of this work the term "structure" means to execute all the activities proposed by the architecture to transform a document into a dataset. These activities are described as primary functionalities in the architecture development process. The Fig. 3 illustrates the architecture overview.

For the development of the architecture we used the ADD method proposed by Bass *et al*. (2012). This method is the most comprehensive and widely used by the industry for more than 15 years in the development of software architectures (Kazman and Cervantes, 2016). The ADD suggests the architecture lifecycle in five sequential activities: 1-Define architectural requirements, 2-Define architectural design, 3-Document architecture and 4-Validate/Deploy architecture.

Regarding the activity "1-Define architectural requirements", Table 1 shows the architecture's main quality attributes obtained after requirement analysis from data availability problems occurring in 561 Brazilian cities' websites (Corrêa *et al*., 2017).

The activity "2-Define architectural design" includes the architectural drivers detailed in the following subsections:

Purpose

The purpose of the architecture is to provide architectural artifacts to implement a collaborative system for structuring information into open data to enable the use and reuse of the data by any and for any purpose.

Quality Attributes

Quality attributes are properties subject to testing and measurement to indicate how well a system meets the needs of stakeholders. The quality attributes specified in the first column of Table 1 were taken into consideration.

Primary Features

These functionalities are the architecture's ability to perform the work to which it was designed, which are:

- Identify the data source in a document
- Extract the tabular content into a dataset
- Catalog the dataset with the addition of metadata to describe its structure
- Load the dataset into the OGD repository

Architecture Interests

They indicate points that a priori cannot be expressed as quality requirements. The architectural interests considered are:

- The architecture is composed of layers for the compartmentalization of its functions and for better division of the components that compose it
- The plug-ins developed to interface with the browser must be compatible with the main products in the market
- Possible new extraction libraries must integrate with the system to ensure high cohesion and low coupling of the system modules
- Architectural implementations and incremental evolutions must be shared with the community in a public and accessible repository by anyone

Table 1: Architecture's main quality attributes obtained after requirement analysis based on Corrêa *et al*. (2017)

Tabular format of the Utility Tree

Quality attribute	Attribute details	Architecturally Significant Requirements - ASRs
Collaboration	Incremental structuring (H, H)	The system allows users to contribute incremental versions based on their own data or from other users
	Collaborative operation (H, H)	The system allows the structuring of data from an undetermined number of sources through the distributed and incremental operation of the community
	Hierarchical moderation (H, H)	The system provides means to combat vandalism, moderated by the user hierarchy recognized by the community's own reputation mechanisms
	Mutual evaluation (H, H)	The system enables mutual evaluation of contributions made by users using simple and objective criteria to build users' reputation
Usability	Browser-based user interface (H, H)	Users have the web browser as their design interface. The browser will introduce the source data source and the work tools through plug-ins developed for the main browsers of the market
	Ease of operation (M, H)	The structuring tools are designed for the majority of the target audience, without knowledge in the area of analysis and data processing
Security	User identification (H, H)	Users are identified solely within the system by means of user and password, maintaining the link of generated artifacts
	Access (M, H)	Users can auto-log into the system without the intervention of other users
Extensibility	New formats (M, M)	The system allows to extend new methods of extraction that allow the treatment of new formats
Non-intervention	Keep generation process (H, H)	Data structuring does not interfere with existing data production processes and will not require data to be compatible with open data
Persistence	Operation track (M, M)	The system records all operations performed by users, as well as the structured and originating data
Presentation	Manual access (H, H)	The system provides manual interfaces for obtaining data, with advanced search capabilities
	Automated access (H, H)	The system provides APIs for access by other tools to be developed and integrated with the system
Interoperability	Machine processing (H, H)	The system should always consider the CSV format to safeguard the data originated from the structuring process

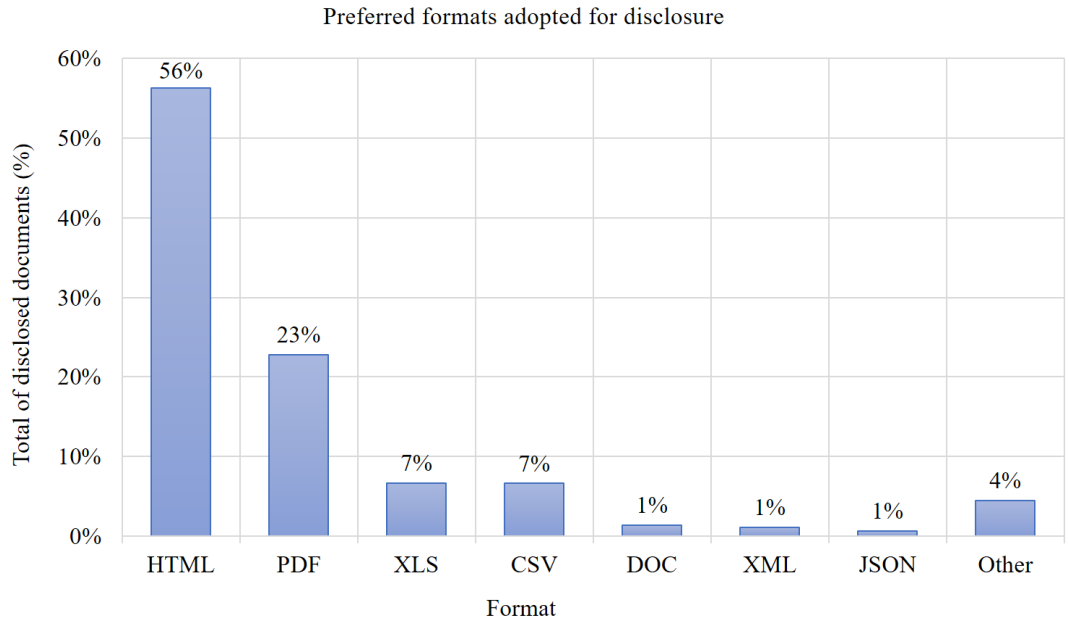

Fig. 1: Main data formats for transparency datasets/documents availability according to Corrêa *et al*. (2017)

Restrictions

Constraints are previously established decisions that make the designer/implementer have little or no control over them. For this architecture, they are the following:

- All libraries, frameworks, development kits, databases and other required elements for the implementation of the architecture must be open source and license-free for any purpose

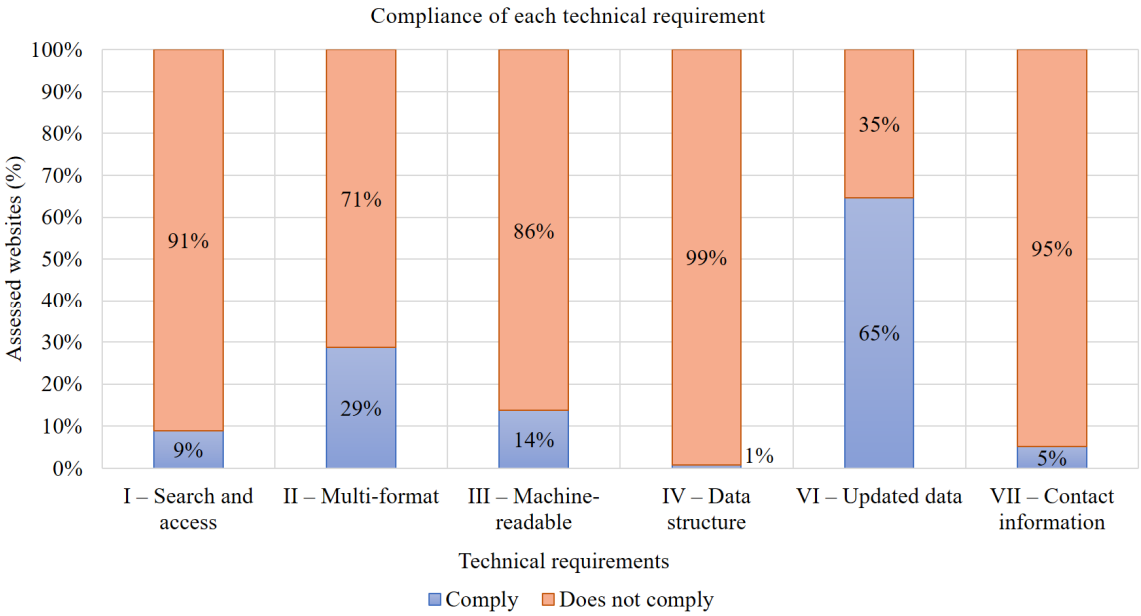

Fig. 2: Technical requirements necessary to perform documents availability according to Corrêa *et al*. (2017)

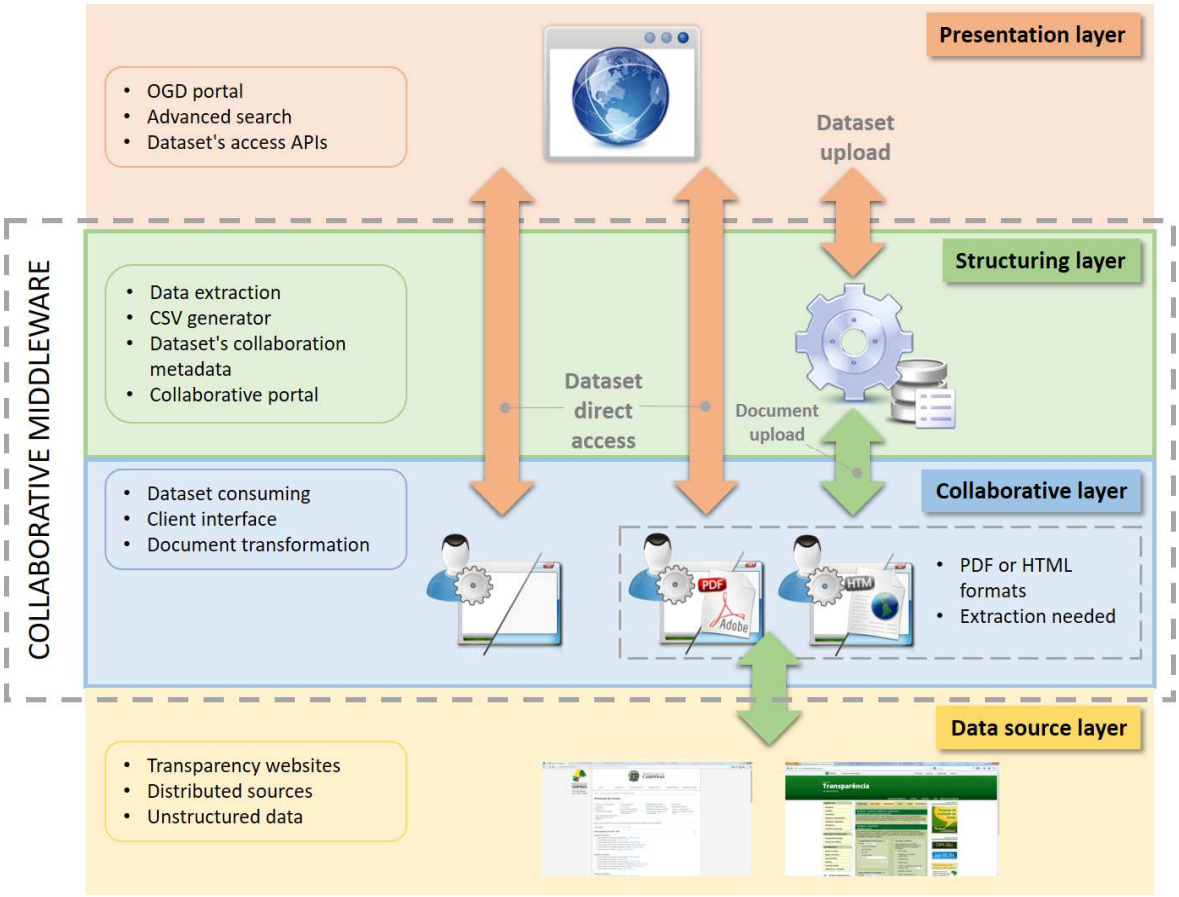

Fig. 3: Overview of the proposed architecture initially proposed in the form of a collaborative-oriented middleware by Correa *et al.* (2014) and Corrêa *et al*. (2015)

In the activity "3-Document architecture" there is the presentation of diagrams, drawings, flows and any other visual artifact that collaborates with presentation and understanding of the architecture by the stakeholders (Clements *et al*., 2010). For the architecture taken into consideration in this study, we reproduce Fig. 3 proposed by Correa *et al.* (2014) and Corrêa *et al*. (2015) as a more comprehensive illustration that exposes its main components and underlying relationships among them.

In Fig. 3 architecture layers are differentiated by the role each one plays. The Collaborative Middleware is located within the Structuring layer and the Collaborative layer. In general, the Structuring layer is responsible for providing data extraction and a repository to record extraction activities to support the collaborative process. The Collaborative layer provides user interface and interaction tools with unstructured data sources.

In the activity "4-Validate/Deploy architecture" there are activities to ensure that the architecture serves to the purposes that it was built. Validation can be done by the designer, his peers and external parties to the process of developing an architecture. For this work, we considered the Architecture Tradeoff Analysis Method (ATAM) (Bass *et al*., 2012; Angelov *et al*., 2008) that has been used for more than a decade for all types of architecture. This method was selected because it is indicated to validate architectures that have not yet been implemented and have many stakeholders involved in their business. In the validation with ATAM a ritual is performed with the participation of people with and without knowledge of the architecture business in order to generate various usage scenarios and the validation of their previously defined quality attributes.

Using Open Data for Digital Forensics

The proposed architecture can be used to the provision of services where structured data is useful. In addition to government activities directed to society in general, an example where it can be applied is the digital forensics analysis which aim to discover patterns of fraudulent activities. In this way, open datasets can be tested against well-known data to provide a basis for comparing methodologies and tools.

A work (Carvalho *et al*., 2016) by one of this paper's authors performed experiments on different open datasets of images. The process of testing against fully available data was essential to reach accuracy of their method and benchmark results with other state-of-the-art techniques.

Another work (Yannikos *et al*., 2014) emphasizes the importance for open datasets for digital forensics education and research. Authors overview the shortcomings when using open datasets and discuss different approaches for each suitable forensic task according the availability of data.

Discussion and Future Works

In this study, we have developed an architecture for structuring documents into open data from distributed repositories, initially arranged in PDF and HTML formats. The architecture was developed by the identification of problems with regard to data availability found in the local government websites in Brazilian municipalities (Corrêa *et al*., 2017) and based on an initial proposal of a collaborative-oriented middleware previously proposed by this paper authors (Correa *et al.,* 2014; Corrêa *et al*., 2015).

Our approach is based on the premise that public information will continue to be disclosed in the "as is" way due to agencies' lack of preparation for open data principles. Thus, it requires time for reengineering of work processes and needs attention for natural evolution guided by society demand. Our proposed software architecture will enable data consistency, since its provision will not interfere in the generation of information, but only in the availability according to the OGD principles for unrestricted access by anyone and for any purpose.

In the architecture context we also considered a collaborative approach to involve stakeholders and open data enthusiasts. This means that its broader use will engage people outside the agencies' boundaries in order to contribute to the structuring process of data. Once structured, data will be made available in data portals complied with OGD for consumption by interested counterparts. The collaborative approach will provide evaluation and classification mechanisms to form a self-sustained and moderated network of contributors.

We also emphasized the use of our proposed architecture for areas that are not rightly directed to society in general. We envisioned the use for digital forensics analysis where structured data is essential for benchmarking against open datasets.

As future work we foresee a development of a software prototype that implements the architecture defined here. With this, it will be possible to verify its feasibility of implementation and operation in real environment.

Acknowledgments

The authors would like to acknowledge Fundação de Amparo à Pesquisa do Estado de São Paulo - Fapesp (Grant Project 2017/12631-6) and Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq (Grant Project 304538/2017-5).

Author's Contributions

Andreiwid Sheffer Correa: Main research conception and design data acquisition.

P.L.P. Correa: Research supervisor analysis and interpretation of data.

F.S.C. Silva: Research supervisor analysis and interpretation of data.

T. Carvalho: Conception and design applied for digital forensics analysis and interpretation of data.

Ethics

We testify that this research paper submitted to the Journal of Computer Science has not been published in whole or in part elsewhere. All references to used data, materials and methodologies essential to this paper were properly included in the text to give to original authors' the right of their work.

References

Angelov, S., J.J.M. Trienekens and P. Grefen, 2008. Towards a method for the evaluation of reference architectures: Experiences from a case. Proceedings of the European Conference on Software Architecture, (CSA' 08), Springer Berlin Heidelberg, pp: 225-240. DOI: 10.1007/978-3-540-88030-1_17

Bass, L., P. Clements and R. Kazman, 2012. Software Architecture in Practice. 3rd Edn., Addison-Wesley Professional, USA.

Bertot, J.C., P.T. Jaeger and J.M. Grimes, 2010. Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies, Gov. Inf. Q., 27: 264-271. DOI: 10.1016/j.giq.2010.03.001

Brickley, D. and R.V. Guha, 2014. RDF Schema 1.1.

Carvalho, T., F.A. Faria, H. Pedrini, R. da S. Torres and A. Rocha, 2016. Illuminant-based transformed spaces for image forensics. IEEE Trans. Inf. Forensics Secur., 11: 720-733. DOI: 10.1109/TIFS.2015.2506548

CGU, 2011. Acesso à Informação Pública: Uma introdução à Lei nº 12.527, de 18 de novembro de 2011, CGU, Brasília.

Clements, P., F. Bachmann, L. Bass, D. Garlan and J. Ivers *et al*., 2010. Documenting Software Architectures: Views and Beyond. 2nd Edn., Pearson Education, Addison-Wesley Professional, ISBN-10: 0132488590, pp: 608.

Corrêa, A.S., E.C. de Paula, P.L.P. Corrêa and F.S.C. Silva, 2017. Transparency and open government data: A wide national assessment of data openness in Brazilian local governments. Transform. Gov. People Process Policy, 11: 58-78. DOI: 10.1108/TG-12-2015-0052

Corrêa, A.S., P.L.P. Corrêa and F.S.C. Silva, 2014. Transparency portals versus open government data: An assessment of openness in brazilian municipalities. Proceedings of the 15th Annual International Conference on Digital Government Research, Jun. 18-21, Aguascalientes, Mexico, pp: 178-185. DOI: 10.1145/2612733.2612760

Corrêa, A.S., P.L.P. Corrêa and F.S.C. Silva, 2015. A collaborative-oriented middleware for structuring information to open government data. Proceedings of the 16th Annual International Conference. Digital Government Research, May 27-30, Phoenix, pp: 43-50. DOI: 10.1145/2757401.2757409

Correa, A.S., P.L.P. Correa, D.L. Silva and F. Soares Correa da Silva, 2014. Really opened government data: A collaborative transparency at sight. Proceedigns of the IEEE International Congress onBig Data (BigData Congress),Jun. 27-Jul. 2, pp: 806-807. DOI: 10.1109/BigData.Congress.2014.131

Dodds, L. and A. Newman, 2015. Modelo de maturidade de dados abertos. Avaliando a publicação e a utilização de dados abertos. Versão traduzida Português BR.

Kazman, R. and H. Cervantes, 2016. Designing Software Architectures: A Practical Approach. 1st Edn., Addison-Wesley Professional, Boston.

Kong, X.Y., 2016. Arquitetura para Publicação de Dados AbertosemSistemas de Governo: O Exemplo do SIASG, Dissertação (mestrado), UFRJ/COPPE.

Machado, A.L. and J.M.P. de Oliveira, 2011. DIGO: An open data architecture for e-government. Proceedings of the 15th IEEE International Enterprise Distributed Object Computing Conference Workshops, Aug. 29-Sept. 2, IEEE Xplore Press, Helsinki, pp: 448-456. DOI: 10.1109/EDOCW.2011.34

Pires, M.T., 2015. Guia de Dados Abertos.

Relly, J.E. and M. Sabharwal, 2009. Perceptions of transparency of government policymaking: A cross-national study. Gov. Inf. Q., 26: 148-157. DOI: 10.1016/j.giq.2008.04.002

Smith, A.W., 2010. Government Online. The Internet Gives Citizens New Paths to Government Services and Information. 1st Edn., Pew Internet and American Life Project, Washington, pp: 44.

Tauberer, J., 2014. The Principles and Practices of Open Government Data. 2nd Edn., Joshua Tauberer, pp: 196.

Yannikos, Y., L. Graner, M. Steinebach and C. Winter, 2014. Data corpora for digital forensics education and research. Adv. Digit. Forensics X, Springer, Berlin, Heidelberg, 2014: pp: 309-325. DOI: 10.1007/978-3-662-44952-3_21