

An account of research on Open Government Data (2007-2012): A Systematic Literature Review

Tharsis Novais¹, João P. de Albuquerque², Gisele S. Craveiro¹

¹School of Arts, Sciences and Humanities, University of São Paulo, Brazil

²Department of Computer Systems, Institute of Mathematics and Computer Science,
University of São Paulo, ICMC-USP, Sao Carlos, Brazil

tharsis@usp.br

jporto@icmc.usp.br

giselesc@usp.br

Abstract: Open Government Data (OGD) is a government policy that is being increasingly adopted by several countries around the world. Despite its growing acceptance, there are still several difficulties both in making OGD available and ensuring it is used effectively by society. A growing body of knowledge has been produced in the last few years in the e-government scientific community to address these problems. This paper conducts a systematic literature re-view on the question of OGD, which is aimed at outlining a knowledge base of the existing scientific literature on OGD produced in the period 2007-2012 that is available in the on-line databases of relevant scientific societies (ACM, AIS, IEEE). In doing so, the paper provides an overview of this knowledge base, and maps out the different research approaches that are being adopted to study OGD, its geographical scope and the application domain they refer to. The results indicate the need of broadening the geographic coverage of OGD studies and also of deepening the discussion of quality criteria for assessing open data.

1 Introduction

Open Government Data (OGD) is growing in importance and becoming widespread throughout the world. Several countries are adopting its policy of public transparency and providing the public with non-confidential information about what occurs in the government, and how public expenditure is handled and distributed, among other information that is important from the standpoint of enabling citizens to acquire a greater degree of social cohesion and togetherness[Na10].

In response to the need to make government data available, the term Open Government Data (OGD) was coined, and the term became popular at the beginning of 2008, after the publication of a set of eight OGD principles drawn up by lawyers in the USA, in December 2007 [Na10]. OGD consists of any data produced by public sectors for anyone to use for any purpose [BMS12] and the principles are designed to ensure that the right to information is observed [Na10].

The government data and information (e.g. on legal, financial, environmental and social matters) have been published in several ways in Web platforms which are specifically made available for the purpose of publishing data. This fact has aroused the interest of virtual communities and research groups in the academic world, as well as the sphere of the government and non-governmental organisations [CGI].

Despite the fact that government bodies are increasingly making data available, it is still a great challenge to ensure that the OGD achieves its full potential, both from the standpoint of the government and of civil society [CSDA13]. Several government bodies have difficulties in making this information available in an effective way; there are problems ranging from technical areas – such as the complete lack of support tools or interoperability with the current systems – to management, organisational and cultural problems arising from the need for changes in the procedures and the means of handling information.

A growing body of scientific studies among the scientific community of the electronic government has centred on the question of OGD in recent years and the purpose of this article is to conduct a comprehensive survey of these studies, or more precisely, on the transparency of the data made available by government bodies. This is undertaken by employing the Systematic Literature Review methodology which uses particular rigorous methods that have been previously documented to identify and critically appraise and synthesise, in a concise way, the studies that have been carried out in a particular area, or on a particular question [DD12].

By embarking on a systematic review that is underpinned by the most important databases of the research community, this article seeks to outline a list of 35 articles with important OGD content and which can serve as a baseline for further analysis. The aim of the article is to provide an overall view of the content published in OGD in the last five years (2007-2012) that is available in the on-line databases of the main relevant scientific societies (ACM, AIS, IEEE).

The remainder of the article is structured in the following way: Section 2 gives a detailed account of the methodology established for the survey of the important content and discusses the underlying research questions for retrieving studies in the area. In Section 3 the results are shown that were obtained from the questions raised. These results are outlined, together with an analysis of the studies found in Section 4. Section 5 summarises the final considerations of this paper.

2 Research Design and Methodology

The methodology employed in this article involves a Systematic Review (SR) and is designed to map out the literature that exists on the community of the electronic government surrounding OGD. This consists of a scientific survey on a particular research area which seeks to bring together the important content on the issue being researched [5,6].

There are several methods that can be employed for the application of the SR. The approach that is adopted in this article was put forward by Dyba and Dingsoyr [DD12] and follows three stages: (i) Planning the review (ii) Conducting the review (iii) Reporting the review. This process model allows crucial issues to be identified that had not been dealt with in a suitable way in research carried out previously. The research was designed before any stage of the chosen protocol was followed. In this article the most significant aspects of the SR are outlined and this is undertaken on account of the large volume of data and information generated from the research that was conducted by means of the protocol that has been described.

The list of the key definitions for planning the SR is as follows:

- *Objective*: to conduct a survey and identify the main primary studies carried out on Open Government Data by classifying the studies according to the research methodology employed, the geographical scope involved, the domain of knowledge used and the way that the data were published, and in accordance with purpose of the study itself.
- *Research Question*: what is the available literature that is concerned with making Open Government Data available?

The resources for searching and selecting studies were defined on the basis of the essential items suggested by Felizardo and Martins [FM10] and are as follows:

- *Identification of the sources*: one of the features of our primary studies is that the studies should be available in an electronic format and Web platform and for this reason the electronic libraries of the Association for Information Systems (AIS), Institute of Electrical and Electronics Engineers (IEEE) and Association for Computing Machinery (ACM) were selected as the main electronic databases. The search engine of Google was also used to assist the research in the databases selected.
- *Language*: English.
- *Keywords*: "Open Data" related to the "Government"; "Open Government Data" related to "Open Data"; "Open Government Data" related to the "e-Government". The terms chosen assist the research study and are aimed at refining the search further because OGD is a relatively new term in academic research and the policies of OGD in electronic format constitute a research topic that is still being defined.
- *Types of studies selected*: Articles – full and short papers – and book chapters.
- *Time frame of the studies*: the most recent approaches were sought in this research study and thus the period was confined to five years (2007-2012) for the selection of studies that meet the needs of this investigation. In this manner, the beginning of this period was chosen in coincidence with the year of the publication of the aforementioned Eight Principles of Open Government Data [Na10].

The selection of studies was carried out in two stages: the Preliminary Selection and the Final Selection. In the Preliminary Selection, searches were conducted in databases that had been previously selected on the basis of the defined criteria. The Final Selection consists of a complete and accurate reading of the retrieved studies based on the preliminary selection.

3 Primary Results

The review was carried out by following the protocol defined earlier, over a period of one month (January, 2013). In obtaining the results of the primary selection, “search strings” were used that were formed of combinations of keywords defined for use in the search engines of the bases selected. Thus a string pattern was formed to have a path to follow, which was adapted for every need of the search tools that were used by the selected database. The string pattern that was used was: “open government data and open data”.

On the basis of the string pattern, the studies were searched in the selected databases. The string was adapted in compliance with the standards of each search engine of the database. By using as search sources the AIS, ACM and IEEE bases, we employed their search engines to obtain the results. Google was also used as a search engine and defined some of the criteria offered by the search engine itself with a view to filtering the research results.

Altogether, 38 studies were found. After excluding the repeated studies that were found in more than one search engines, this was reduced to 35, and this list is available on line (www.gpopai.usp.br/listSR-OGD). Once these duplicated studies had been removed, the Abstracts of each article were read and a further 5 studies had to be excluded because they did not meet the requirements of the conditions set out for the research. When all the studies had been read, a total of 30 studies were assigned to forming a part of the corpus for retrieving the data and information. The articles retrieved from the databases were analysed by means of inclusion and exclusion criteria, defined by the aforementioned research question.

4 Analysis of the results

In this Section, there will be a discussion and summary of the results obtained from the research questions of this survey.

Setting out from the study and analysis of the results obtained from the selected works, several approaches can be employed with regard to the material contained. In Fig.1, this can be observed more clearly through the classification of the articles that correspond to the methodology employed for the research. We can see that most of the articles are concentrated on Case Study and Design Science research methodologies, with 11 articles for each. Together these two methodologies represent a total of 74% of the articles surveyed. It is very common in OGD to find articles that deal with a case study of existing OGD initiatives. Papers that employ a Design Science Research approach, on the other hand, usually present some kind of tool and/or framework to assist in the understanding or recovery of open data supplied by the platforms of electronic government.

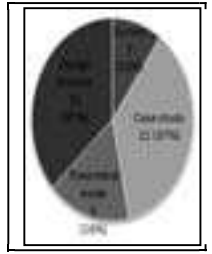


Fig. 1: Classification of articles in accordance with the methodology employed.

Fig. 2 presents the results of a classification of the selected papers according to the discussion of the 8 OGD principles. Thus it can be seen in Fig. 2 that only 4 papers (14% of the studies) in fact discuss all the 8 OGD principles, whilst 6 papers (20%) discuss some of the principles. On the other hand, 4 papers (24%) discuss specific principles, i.e. they discuss the open data principles of a particular country or region (case study). The number of studies that do not discuss any of the principles shown in the graph is compromising since there is a total of 50% or in other words, 15 studies out of the 30 investigated with relevant content, do not discuss the OGD principles.

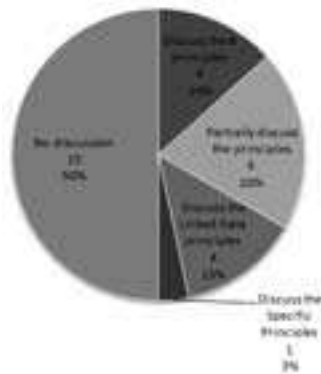


Fig. 2: Classification of studies in accordance with the discussion of the OGD principles.

The same applies to the articles that address the question of the Linked Data principles (only 4, representing 13% of the total figure) since not all of them discuss and/or make use of the Linked Data as solutions to generate new information or adopt this type of content in the tools and frameworks or discuss these principles.

We also carried out a classification of the geographical distribution of the countries that were cited in the studies examined, as can be seen in Table 1. The countries cited in the studies are mentioned in the following ways: (i) as a model to be cited with regard to OGD, (ii) as a place where measures are being taken with regard to OGD. It is clearly apparent that the US is the country where these studies are most widely used.

Most of the studies refer to the US as an example of a country that adopts policies for publishing open data while always offering a tool for the consumption and publication of these data. Although Singapore also forms a part of the best policies for publishing open data, it was only found in one of the classified studies.

In Fig. 3, there is now a classification of studies investigated in the light of the knowledge domain employed in the research. Not all the studies had a specific knowledge domain in a total of 11 studies. In Figure 3, a more salient kind of knowledge domain can be seen which is the "Public budget", with 12 studies that use these data. It was already expected that this would be the most prevalent knowledge domain in the studies because when OGD policies are referred to, one almost always imagines expenditure from public coffers which is what most concerns citizens who want to have the right to know (as tax payers) how their public funds are being spent. However, although the public budget is the most widely used category, it is not the only one. There are also several topics that are found in the other studies and 7 of them make use of, or deal with, data other than the public budget.

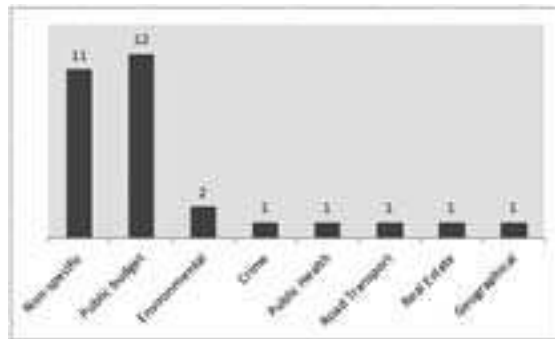


Fig. 3: Classification of the studies in accordance with the knowledge domain.

In Fig. 4, the information was classified in a way that shows how the studies made the data they used or the artifacts (i.e. tools/software) they developed available, so as to assist the OGD research community. In all the literature that was reviewed, 64% (a total of 19 studies) are not classified in this category because they did not produce any artifact or result in any data being made available.

However, 11 studies produced some kind of artifact or else left data available for observation. 7 out of 11 studies left the data that were used, (or the produced artifacts), available externally through a web address and in a way that allowed them to be used. The other 4 studies only made the information available on their own paper, which could make it difficult to continue the research because the data and tools are not available externally in a way that can make them easy to use.



Fig. 4: Means of making available the artifacts/data produced in the studies.

Finally, it is worth noticing that a limitation of this study lies in that the databases used for the searches, since they may let out important outlets that are not directly affiliated to the aforementioned scientific societies. Possible examples are journals with relevant OGD production that are not listed in the employed databases. Future work may apply the methodology composed for this work to other databases so as to expand the coverage of the results.

5 Final Considerations

This article shows the results of a systematic review of the publications on Open Government Data that are indexed in the on-line databases of the scientific societies related to Electronic Government (ACM, AIS, IEEE). Relying upon a rigorous and repeatable search procedure, 35 articles were identified as having relevant content and form a good summary of the research conducted on OGD, which can provide a baseline to help interested researchers in studying this topic.

The results that were achieved show the growth of OGD policies throughout the world and also the increasing interest of the academic community in focusing on this issue. However, this research has also provided evidence of gaps in the existing literature which need to be filled by the research community. The two main gaps that we have discovered are as follows.

The first concerns the geographical scope of the OGD research studies. The literature that exists is still largely concentrated on studies of the U.S. followed by continental Europe and the United Kingdom which is natural in view of the fact that these countries are pioneers in this public policy area. However, there have been important experiences in other countries (such as Singapore and Brazil) which are under represented in the current research. It is thus important for the scientific community to take steps towards broadening and deepening its coverage of OGD studies.

The second serious gap that has been identified concerns the study of the quality of the open data made available. In this regard, the eight principles of open data and the concept of Linked Data constitute key reference-points that can be used to assess the quality of the data made available by the government. Despite this, this review has shown that, when making an appraisal of the quality of OGD, many articles fail to

discuss any of the eight principles or any other framework. In addition, there are still very few studies that discuss the Linked Data.

This is a thus significant factor in guiding future research studies of OGD, given the fact that providing data is not the same as supplying information. Releasing a large number of datasets with dozens of numbers contained in them does not mean anything unless there has been a supply of knowledge that has been made available beforehand. The government must give incentives to its different bodies to provide services that make data available that can be reused without any difficulty and which have portable formats (i.e. that can be consumed, regardless of their system or architecture) and which can enable new information to be generated.

Hence, it is necessary to broaden and streamline the architectures and analyses available in the literature about how to improve the provision of OGD and the tools that are developed to generate information that can be easily understood by the general public. In this way, the scientific community can assist in a positive way in allowing OGD to achieve its full potential for the benefit of society.

Acknowledgment

João P. de Albuquerque would like to thank FAPESP (process 2012/18675-1) for providing partial funding.

References

- [BMS12] Bao, L., Mao, F., Sheng, Y.: Analysis the Present Situation of Chinese Electronic Government Based on Fundamental Data. In: Eleventh Wuhan International Conference on e-Business, Paper 97. AISeL, Wuhan (2012)
- [CGI] CGI.BR.: Melhorando o Acesso ao Governo com o Melhor uso da Web, <http://www.w3c.br/divulgacao/pdf/gov-web.pdf>
- [CSDA13] Craveiro, G. da S., Santana, M. T. De, Albuquerque, J. P. de.: Assessing Open Government Budgetary Data in Brazil. In: The Seventh International Conference on Digital Society, pp.20-27, ICDS (2013)
- [DD12] Dyba, T., Dingsoyr, T.: Strength of Evidence in Systematic Reviews in Software Engineering. In: 34th International Conference Software Engineering, Canada (2012)
- [FM10] Felizardo, K., Martins, R.: Engenharia de Software Experimental: Uma Revisão Sistemática. In: XIII Congresso Iberoamericano en Software Engineering, Cuenca (2010)
- [Na10] Nam, T.: The Wisdom of Crowds in Government 2.0: Information Paradigm Evolution toward Wiki-Government. In: Americas Conference on Information Systems 2010 Proceedings, Paper 337. AISeL (2010)