Collaborative Construction of an Open Official Gazette

Gisele S. Craveiro, Jose P. Alcazar, and Andres M.R. Martano(✉)

School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, Brazil
andres@inventati.org

Abstract. Given the potential use of open data and the obstacles for implementing Open Government Data (OGD) initiatives, this paper aims at describing the strategies adopted for preparing the implementation of an open Official Gazette at the municipal level. It is important to emphasize the potential value of the Official Gazette as a source of information, since it is perhaps the most detailed and comprehensive report the society can have on government daily activities. However, the data are mostly unstructured, and this fact, combined with the size of the database, makes any attempt to analyze it a non-trivial matter. Publishing the Official Gazette as OGD certainly does not address all the problems related to its use, but hopefully barriers can be overcome to allow more groups to make use of it. In this paper, three research methods are combined; a bibliographical review, documentary research, and direct observation. This paper describes the strategies and activities put into effect by a public body and an academic group in preparing the implementation of the open Official Gazette. It also analyses the outcomes of these strategies and activities by examining the tool implemented, the traffic and the reported uses of the Open Gazette. The paper concludes by reflecting on the main challenges that are raised in implementing open data initiatives at a local level in a developing country, and proposing an agenda for future research.

Keywords: Open government data · Official gazette · Coproduction

1 Introduction

The technological advances are changing the way the citizens have access to public information. Not only the commoditization of computers and other information and communication technologies, but also the access to public information is more and more disseminated in open format, allowing greater reuse driving innovation in the public sphere. This change is happening on the way we create and exchange knowledge and culture, as well as how we participate in civil life [1].

One of the main sources of public information are the official gazettes, since it is through them that the official acts are not only made public, but are also considered in force. That is, they are only applicable from the moment they are

© Springer-Verlag GmbH Germany 2016
A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 111–128, 2016.
DOI: 10.1007/978-3-662-53416-8_7

published, or in cases specified on the acts themselves, in periods counted from these publication dates. For instance, in Brazil, this practice goes back to 1808 and, since then, the daily publications seek to ensure universal access of citizens to the public acts, as well as their historical record.

Naturally, this type of publication suffered the impact of the arrival of new information and communication technologies, and the official gazettes are now made available both in printed paper and in digital media. Usually, the digital version is offered in a PDF format, which has inhibited or even prevented the automatized reuse of the published information. Therefore, there is an excellent opportunity for publishing this documents in an open format and expand the current notion of universalization of access existing in public administration.

There also are great challenges to provide the availability and consumption of such information. The main ones are related to the unstructured or in some cases semi-structured nature of information, as well as the lack of standardization and controlled vocabulary on nomenclature used in public acts. Also, consultations and analyses on the daily volume of several years of publication require strategies for organizing and making the official gazette available in open format to all citizens.

This paper presents the process of developing an open 'official gazette' in Brazil. Although there is a growing interest in how to make governments more transparent, and in particular how to make government data open to a broader public, there is a lack in the scientific literature describing the effective development of such iniciatives. We aim to give contributions to the discussion about process and the system architecture from a real experience implemented in the local government. This study describe the process of opening the Official Gazette in the city of São Paulo, performed via partnership between the municipal public administration and a group of researchers. The methodology for constructing this initiative, named Diário Livre (Free[1] Gazette) is described, both in collecting the demands and expectations for the project and also in its technical implementation. Its initial impacts will be discussed, stating some benefits noticed from some of its consumers and the challenges faced during its implementation, as well as its maintenance. Finally, the paper will be concluded by presenting the future actions and final considerations.

2 Background

2.1 Official Gazette

Official gazettes are government newspapers for disseminating information, created to give publicity to measures taken by the government. They can have several names and have a wide coverage. In Brazil, the origins of the official gazettes go back to the time when the Portuguese Court was transferred to this country. It was in 1808, when the Royal Printing Press was established by

[1] "free" as in freedom.

decree in Rio de Janeiro and was granted exclusive rights to print all the official legislation and administrative measures from the government. Currently, the responsibility for publishing the Official Gazette is of the National Press, a body that has close ties with the Presidential Office of the Republic. There is a decree determining the scope of what subjects can be published in the Federal Official Gazette (DOU) [2].

In the case of the municipality of São Paulo, the Diário Oficial Municipal (DOM) (Official Gazette of the City) is under the responsibility of the Secretariat of Planning, a secretariat in the So Paulo City Hall, and published by the Official Printing House of the State of São Paulo.

In accordance with the requirements of the DOM site [3], the publication is divided into 7 sections and covers: decisions about processing, authorization and contracts, exemptions, nominations, appointments, competitions, expenses statements, responsibility and salary reports, balance sheets and more.

Thus, it should be stressed that the material supplied by DOM is very important to ensure transparency and government accountability, since it displays information about servers, public expenditure and decisions that affect the community as a whole.

2.2 Open Government Data

In a context where there are increasing talks on government innovation being made together with the society and not only for the society, the open government data is the raw material in co-creation processes between government and the civil society. The open government data (OGD) are described by eight principles [4]. These principles, together with the 5-star model [5], seek to allow new uses for the data produced. Therefore, their publication is not usually seen as an end per se, but as a means for possibly producing a positive impact on society [6].

It must also follow the principles found in Open Definition [7] "open data are data that can be freely used, reused and redistributed by any person". This involves the online publication and sharing of information in open, machine-readable formats, that can be freely and automatically reused by the society. The former will ease transparency and democratic control, citizen empowerment, innovation, improvement in government services and the discovery of new things from the analysis and mining of data [8].

There are many political benefits from open government data. These include the following: an increase in transparency, an upgrading of public services, an ability to combat corruption, an opportunity for innovation and an increase in government efficiency. This paves the way for greater involvement of the people, and the creation of new markets which can make use of the available data and generally improve decision-making, as discussed by [6,9,10]. However, although this trend has gathered momentum in the last few years and influenced governments throughout the world (by encouraging the adoption of a large number of similar initiatives), the real effects are still relatively unknown [10,11]. However, preliminary studies suggest that the effects may be negative as well as positive, contrary to the more optimistic expectations [12].

As discussed earlier, a wide range of factors including technical constraints, can mean that only special sectors in society can make use of OGD. This asymmetry of access can thus lead to greater inequality rather than reduce it [12–14]. Given this situation, it is clear that any decision to implement an information system that can provide open data, may also influence the policy making that can bring about its appropriation.

As a result, some of the measures recommended for the publication of DGA, which seek to encourage their use for positive ends, entail a greater involvement of society in ensuring that those groups that lack technical qualifications, are assisted and supported in their attempts to make use of the data. With regard to this, [10] stresses the importance of empowering those that are already carrying out this method of using DGA, as well as connecting users, developers and public managers, since it is generally found that managers are unable to visualize the best ways of employing the data. This interaction between groups is essential in deciding which data should be open and in what way it should be published. But, regrettably this kind of interaction is not the rule, and can, for example, enable many managers to publish data in a way that they believe to be open but which, in reality, does not allow a series of uses.

The three documents address the question of political feasibility and show its importance. The related recommendations involve working with the prevailing culture of the organization by employing prototype projects for the experiment carried out with open data and to demonstrate its benefits. With regard to the licences for the data, there is a general consensus that they should be open. This belief is linked to the fact that the three documents also support the idea of exploiting these data for commercial purposes – something which can be constrained by the more restrictive kinds of licences.

Two of the documents seem evidently to support a more participative opening process. This is made clear in several points. First of all, in arguing that the choice of data should be open, they state that this is the responsibility of a consultative group that is outside the control of the public administration. Following this, they suggest forming a catalogue of data which can enable the community to know what kind of data exist so that they can have a greater sense of proprietorship with regard to the prioritization of the opening.

Another key factor which is also an advantage with regard to participation, is the support provided by the use of open formats and free software, which allows a degree of collaboration in the development of the tools that are used for a greater interaction with, and appropriation of, the data made available. In addition, and also linked to participation, the three documents agree that some kind of partnership should be formed with the intermediaries to assist in the process as a whole, or in some particular phase.

Finally, with particular regard to the data consumption, the documents also discuss the idea of fostering this through events (such as hackathons or courses) or partnerships. It should thus be noted that, on the whole, there are no wide divergences between these three documents but only some differences of standpoint and that they can operate in a complementary way.

2.3 Related Work

The open government data initiatives are generally published as numerical data and are not strictly documents like those that constitute the Official Gazette. Although there are some works in the literature that tackle the question of machine readability applied to the Official Gazette [15–17], there are hardly any studies of this kind and they fail to address the question in its socio-technical totality. As a confirmation of the importance of this case study, not many studies were found about how to publish Public Gazettes as open data. One of the few studies that was found commented on an initiative concerning open data in the Philippines. This study stated that the first data to be published by the initiative consisted in digital versions of the Public Gazette for that country. However, the quality of its metadata has been criticized [18].

In another article, the authors had to examine the Brazilian Federal Gazette in order to make semantic annotations on the articles and link them to each other [17]. One of the results obtained was the ability to establish which acts had been annulled or superseded by others, although the only domain addressed was confined to the Treasury. The Brazilian Senate LexML was used as the basis for identifying the acts in the text of the articles and it seems this has greatly assisted the procedure. However, it was very hard to identify the signatures on the acts, since there is no defined vocabulary for handling the names of the people concerned. The authors did not make it clear what issues had been found while extracting the text from the PDF files or whether this process may have had an effect on the results. However, they state that the text in the PDF was organized in several columns which must have caused a great deal of difficulty in their extraction.

Finally, the last article found discussed the publication of data (through SPARQL) regarding the laws of Chile [16]. But, although the architecture chosen for supplying and documenting data has been outlined in detail, it was not clear which format was initially used for the database.

The award-winning initiative Federal Register 2.0 [15] is also very interesting, which seeks to convert printed material into machine-readable XML data. Its goal is "to make the *Federal Register* more searchable, more accessible, easier to digest, and easier to share with people and information systems". Despite being mentioned in some works as an example to be followed [19,20], we were not able to find details on the project and implementation in scientific literature.

3 Methodology

This project was developed through a partnership between the public sector and the researchers, and used several methodologies to support its development. As well as the bibliographical review on official gazettes and their availability on the internet for seeking related experiences, a documental and experimental research was also performed.

The methodology employed in this work is inspired by the Action Research (AR) design. This choice was made because AR predicts and supports the intervention in the process, something which has occurred in this study because of the partnerships. According to [21], this strengthened approach includes a period for the establishment of the research environment and then the cyclical iteration of 5 phases: **Diagnostic**: formulation of theories about the causes of organizational problems; **Action planning**: the definition of the measures to be taken to tackle the problems based on the theoretical framework and register of projected aims; **Action**: implementation of planned actions, either personally or through third parties; **Assessment**: if they are successful, an assessment of whether the changes can really be of value and if there is a failure, if the framework and theories can be adapted; **Specific Learning**: structuring of the acquired knowledge, whether or not it has been a successful experiment.

In the domain of Information Systems, methodologies of this type have been employed by [21,22], both claiming to have obtained good results. The first study supports the view that, within the area of this research, the AR methodology is justified, especially when human organizations interact with information systems owing to the complexity of this interaction. However, neither of the studies devotes much space to the political context or the power relations involved in the implementation of these systems, which are factors that are fully discussed by [23]. In the opinion of this author, the AR should not just seek to obtain better results that are based on practice, but also act in a way that can reveal the power relations which can bring about subversion. In this way, the methodology can empower the participation of the process through a real inclusion.

Bearing in mind what has been outlined in the previous paragraph, this project has sought to employ an AR methodology. In the first meetings between the academic group and the policymakers, the initial scope of the project was set out. A team was appointed to carry it out by forming a partnership. In the course of several meetings, this took care of the planning cycles, action and reappraisals. Although there were some smaller cycles, in general terms the project took place as follows: (a) a stage to conduct a survey of the required conditions, (b) the construction of a prototype and its launching, and finally (c) a stage for gathering the impressions of the users and players involved.

It should be stressed that, owing to the inherent features of the methodology employed, these stages were not followed rigidly and there were some interactions between them. It can be said that these stages and the process as a whole, were mainly influenced by: discussions held during the periodic meetings of the team; the internal context of the municipal authority (coordination with secretaries, disputes between members of the inner group, availability of the data, etc.); capacities of the team (availability of time, technical knowledge); expectations of members of the team; contributions made during the first public event.

In addition, as a means of having a better knowledge of the context, the following results were achieved which could act as guidelines for the activities of the project and to ensure some degree of participation: an in-person questionnaire at the first event; an online questionnaire at the tools site which was during the

whole process; a collection of statistics for access to the tool; interviews with the managers responsible for the database; a second public event for the launching of the tool and gathering of impressions.

For the survey of the current scenario and demands for the project, over 10 meetings were held with public managers from at least three secretariats related to the collection, organization and availability of public information gathered in the official gazette. As well as the continuous alignment of expectations with the execution of the implementation between public administration and researchers, the demand for opening and expansion of social participation in the construction of this software artifact has led to the existence of two events open to the public.

The first event was held on the dissemination of the project for collecting the demands of the citizenship, and the second public event was performed both for accountability purposes, as well as for the broad dissemination to reach the social actors who are interested in and have the means to consume and reuse the volume of data now available in open format. The information obtained in these meetings (restricted and open) were gathered through questionnaires, semi-structured interviews and observation records.

All information collected supported the design of the requirements and later project and implementation of the tool, subdivided in extraction, transformation and load modules. The technological development process first lead to standardization procedures of files made available by the public management related to eleven years of daily publications. After the initial treatment, information was organized on a base for the possibility of later creating indexes for the individual articles. On this base, some functionalities were implemented, such as search tools and different forms of publication. Due to the evolution of the negotiation process involved in obtaining data from public power, the implementation of the automatic extraction stage for daily updating of data can only be completely integrated to the system at the end of its implementation. Aiming to increase the replicability of the process, ease its adaptation and reduce costs, all software used in this project are open software.

The resulting tool, named Diário Livre (Free Gazette) was delivered as a proof of concept to the public management, and was made available from the research group infrastructure. Its official launching happened in October, 2014, and since then, the service is provided uninterruptedly, automatically collecting data from the public power and making them available through a web application.

After its launching in a public event and in the media, the number of accesses has been monitored, as well as the integration performed through e-mails or events where it is disseminated. It is important to mention that one of the main public consumers consists in the civil servants and thus, a message disseminating the tool and requesting its evaluation was sent to over 200 thousand government employees in the city of São Paulo.

An important qualification needs to be made about what has been outlined here. This study was carried out through a partnership with an administrative public body. This made it possible to conduct an in-depth analysis of the internal public mechanisms and automatically add a group to those responsible for

making decisions about the project – the policymakers. On the other hand, this partnership imposed constraints on attempts to participate with other groups, either because this was the will of the policymakers involved or on account of the restrictions that were self-imposed. These factors added to the difficulty of employing the AR in the form supported by [23]. Another important point that needs to be underlined is the fact that the authors were only able to employ the methodology by intervening in the process – something anticipated and supported by the AR methodology. However, as stated earlier, there was no complete control of the procedure.

4 Development

In this section, the project development stages are presented in greater detail. First it describes the scenario found when the project was started, presenting which data was available to be worked on. It then lists which were the requirements surveyed for an availability that would improve the initial scenario. It also presents and justifies the architecture adopted for the availability of data. And finally, it describes how the architecture was implemented to meet the requirements of the project.

4.1 Scenario Found

The Official Gazette from the City of São Paulo has a curious flow. Its first stage is open, with information stored on a file in the ".txt" format. However, for its printed publication, through the Official Printing House, or even for online availability with legal value, a "PDF" file is generated. Due to its nature, this file format is closed, that is, it does not allow its information to be easily copied, handled or researched. The flow from the generation of the governmental information until its availability to the public is detailed below.

The Municipal Secretariat of Management (Secretaria Municipal de Gesto - SMG) is responsible for publishing the Official Gazette at the City Hall of São Paulo. This secretariat is responsible for hiring the Official Printing House (Imprensa Oficial - IO), the same company responsible for publishing the State Official Gazette, to publish the Municipal Official Gazette (Diário Oficial Municipal - DOM). Figure 1 represents the initial flow for publishing the DOM before the tool for publishing open data.

In order to a member of staff from any department in the city hall to be able to publish anything in the municipal Official Gazette following the guidelines from the IO itself, the text has to be written and saved as TXT. The name of the saved TXT file must contain a code, known as *retranca*, which represents two metadata: the content of the article and which public body has written it. This file is then sent to *Pubnet*, the IO system for collecting articles. Images, named *calhau*, follow a different process, being saved as PDF and not as TXT before being sent to publication [24].

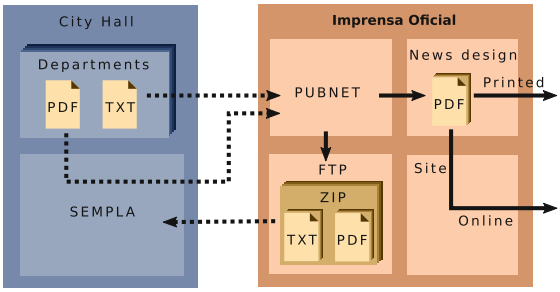

Fig. 1. The initial flow of data.

Once sent to IO, the material is saved and fully available, through FTP protocol, back at the city hall. This material, constituted by one ZIP file per day, is important, since it allows later checking if, for instance, the city hall decides there is an error made by the IO in publishing the Official Gazette. The material is also sent for editing, where it is formatted to be published. The latter is made on paper (printed newspapers) and online, making the PDFs available at the IO website.

The PDFs keep the same visual structure of the printed newspaper, very different from the structure of the initial TXTs. Such fact hinders the automatic extraction of text, but eases the comparison between the printed and digital version.

The form for textual search on the site, which indexes DOMs since 2005, allows filtering by publishing date, but does not allow filtering by publisher or published content, metadata available in the initial TXTs.

Even though the full database form the Municipal Official Gazettes is not currently available for download at the IO website, it is possible to write programs to automate the extraction of the PDFs, page by page. From them, the text from the articles can be extracted. However, such process is not precise, since the PDFs are not structured. Such procedure is performed, for instance, by companies that provide services of mining data on the extracted base.

4.2 Requirement Analysis

The requirements of the tool were raised through meetings with the managers involved in the partnership, in events with the community, from literature on open government data and from the answers to the questionnaires. The main requirements surveyed are described below, divided into three categories.

The first category consists in the requirements that are provided by the official site and were important to be kept in the new tool: **Daily updates:** the DOM is published every day and the site must be updated with new data with the same frequency; **A fast search:** even after being processed, the database included approximately 10 gigabytes of texts. An efficient tool for textual search was required in order to obtain the results and display them on the site in few

seconds; **User notifications:** allow the users to be notified when there are new publications containing certain key words. Therefore, the user can be notified when, for instance, his name or any subject he is interested in are mentioned in a new article in the Official Gazette.

The second category consists in the requirements that are provided by the official site and should be improved in the new tool: **Search filters:** it was important to allow the data to be filtered through the available metadata, such as, for instance, publisher and date of publication.

And finally, the third category includes the requirements that are inexistent in the official website, but should be implemented in the new tool: **Text visualization:** many users complained about the difficulty of copying texts from the PDFs published by IO or about the loss of formatting when pasting copied texts. The site must return the articles as text in a common HTML page; **Access via API:** it was important for the site to offer some kind of API to allow easy access to the data through other applications; **Full database availability:** this allowed the use of the database as a whole, either for research or for applications that demand a greater control of data; **Unique URLs for articles:** it was important for each article to be accessible through a unique URL, so that it could be easily quoted.

4.3 Architecture

The *Diário Livre* (DL) tool was designed, implemented and launched on the basis of the requirements that were expressed. From the standpoint of a common user, it basically consists in a site where it is possible to search for words in the DOM from 2003 to the present day. Some filtering can be applied for the publisher, article content and date. The standard output format is a common page (HTML), although it is also possible to search for and visualize the articles in other formats.

The architecture adopted for the new system is represented at Fig. 2.

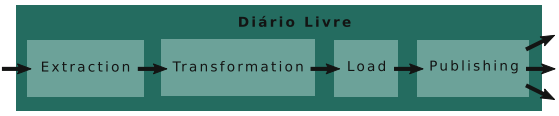

Fig. 2. Diário Livre architecture in layers and information flow.

On the extraction layer, the data are forwarded by the public management and inserted in the machine where the system is hosted in order to allow the processing of data. Due to the internal security policy at the city hall and agreement issues between it and IO, the system cannot collect data within the city hall network. Instead, the city hall staff need to forward the data to the publication system. And since, in a first moment, this transfer needed to be manually

performed by the city hall staff themselves, the tool used in this stage needed to be usable by non-developers.

During transformation, the data are standardized and prepared for indexation and publication. The data provided by the city hall are compressed and include several file formats. They need to be treated and standardized, identifying the encoding of the text file contents and the metadata implicit in the name of these files, and only then, it is possible to index the data.

On the third stage, loading, the data are indexed to allow searches. Due to the size and characteristics of the database, it is necessary to have a tool that is able to index large amounts of text, perform textual searches in a timely manner and filter by metadata.

On the fourth stage, the data are finally published in several formats. These formats encompass the conventional web viewing, full download from the base and access through API, seeking to maximize the number of possible uses. The URLs used in the conventional web viewing or in the API allow individual access to the articles, and not by page, as it happens on the official IO system.

4.4 Implementation

Figure 3 represents the current flow of publication including DL. The two upper rectangles represent the flow between the city hall and IO prior to DL, and it continues to produce the official publication.

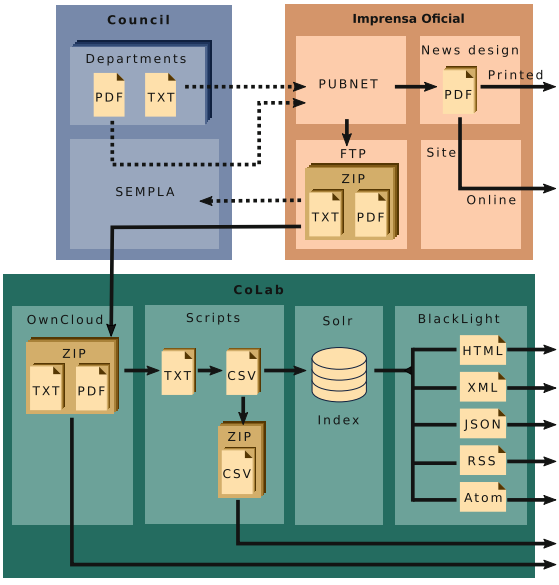

Fig. 3. The flow of data today.

The data, compressed in ZIPs, are sent by the city hall to the machine where the system is hosted. The TXT files are extracted, standardized and organized in CSVs, where each column has one metadata (date, content type and publishing body) and the whole content of the articles on the last one. These files are then indexed by the Solr [25] tool, allowing textual search and filtering by any of the metadata. Apache Solr, as presented in [26], is a NoSQL technology, that is, non-relational, which is optimized to resolve a specific class of problems for specific types of data. Solr is scalable and optimized for large volumes of data (data centralized in text) and returns results classified by relevance. The use of this tool allowed the execution of searches in a timely manner.

Once indexed, the data are published in several formats (HTML, JSON, XML, RSS and Atom) via the BlackLight [27] tool, which is used as a web interface to Solr and as an API for automated access to data.

As well as individual articles, two versions of the database are made available. One identical as the one forwarded to the city hall, containing all the files, but without treatment or standardization, and the other containing only the CSVs already treated. The first database, when decompressed, has approximately 50 GB, and the second, approximately 15 GB.

5 Results and Discussion

The difficulties in handling the official gazette in PDF has led to the offering of specialized services from several companies, which performed web scraping of data to sell solutions to clients. However, as well as being restricted to a privileged public, such solution is still prone to errors.

This way, both the citizens who wish to research basic information or even categorize and handle large amounts of data, as well as civil servants who need to monitor specific administrative acts, (such as, for instance, appointments, exonerations, waivers of public biddings, etc.) have difficulty in operating the traditional PDF version of the publication, since it is difficult to locate considerable part of the publications, and its handling is extremely impaired.

In face of the demand for democratization of the access and the need to automate searches for establishing internal control mechanisms, an experimental version of the Official Gazette in open format was developed, named Diário Livre. The prototype was developed and is working since October 2014, being used by citizens, civil society organizations and civil servants. At the time when this paper was written, the resulting site was available at: http://devcolab.each.usp.br/do

To a great extent, Diário Livre solves many of the difficulties that were previously presented. Since it receives information in open format, it makes such data available in the same manner. It is then possible to copy, extract and handle information from the Official Gazette, and now they can be read and processed by machines. Moreover, Diário Livre makes it possible to perform more qualified searches. Through it, information can be researched by Entity (secretariat of finance, hospital authorities, etc.), by Department (offices, directorates, etc.), by

Content type (dispatches, contests, bids, civil servants, etc.), and by Publishing Date. When accessing the results, it is also possible to classify them by relevance and data, automating the searches even further. It is also possible to download the entire database used in Diário Livre with information since 2003.

Considering the period of daily publications from 2003 to 2014, more than 1.4 million files in text format were obtained from the public power, containing a single article (public act) in each of them. Furthermore, approximately 61 thousand files in PDF and over 14 thousand files in .doc format were also received. In terms of volume of data, the files in text format totaled 13 GB and the files in the other non-text formats are a total of 10 GB. By analyzing the non-text files, it could be noticed that most of them contained images and layout proofs, and that the relevant content was in the text files, and for this reason, only these were considered to be part of the database. Nonetheless, the tool provides full disclosure of all files received from the public power, in all formats, through the bulk download. The tool also makes the version treated in the database used for indexation fully available through the bulk download.

The tool automatically collects the information generated on public acts from the municipal public power in São Paulo (executive, legislative and audit court), which guarantees the offer of updated information. This means that the average publication is of approximately 500 new articles per day and 127 thousand articles per year.

The published data by the Diário Livre were licensed as Creative Commons 4.0, (which is an open license). They are made available in machine-readable and non-proprietary formats. Each article is also made available in the formats stated above, and has its own URI. Thus, it can be said that according to [5], they deserve to be awarded 4 stars.

Functionalities aiming at the common user already existing in the official conventional platform, such as filter by data and textual search tool, are also offered in Diário Livre. Diário Livre also makes it simpler and more convenient to mark and copy texts, and allows filtering by categories and full access to raw files, making it easy to search the mass of data from a series of over ten years of publication, characteristics that are not observed in the conventional version. One of the functionalities offered, the automatic notification of terms, allows any person to receive, in a simple and free form, a service that is currently offered upon the payment of a subscription to the conventional official gazette.

But the characteristics that have no parallel in the official version are related to the potential reuse promoted by the automated consumption of these data. With the exporting of data through APIs, it is expected that Diário Livre eases or makes solutions for data mining viable, as well as new applications of social and economic interest in the city of São Paulo.

5.1 Dissemination And Repercussions

The system had two main dissemination moments. On the first moment, the launching, as well as a presence-based event, news was published on the websites belonging to the city hall and the university linked to the project. Each of these

pieces of news was republished at least once by groups that were not directly involved in the project. During the same period, a member of the team was invited for an interview in a private radio broadcasted throughout the state. On the second moment, a dissemination e-mail was sent to all employees at the city hall. There was also an announcement about the system in the DOM itself.

A free software called Piwik was employed to monitor access to the site. Figure 4 shows the number of visitors per day based on the data collected. Two peaks can be seen that are related to the two moments aforementioned. The first moment occurred at the end of October and corresponded to the launching of the system. The second took place in the middle of December, corresponding to the sending of the dissemination e-mail and the announcement in DOM. The frequency fluctuation on the graphic is due to fewer visitors accessing the site on the weekend.

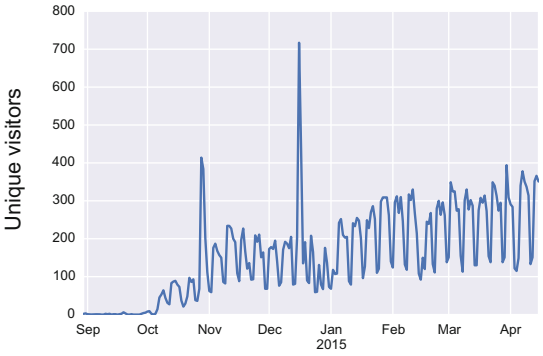

Fig. 4. Number of visitors per day.

Automatic visits, such as those made by search engines, are not included in the charts. It is interesting to note that this increase has started to take place even before the launching of the system. Considering all types of access (automated or not), on average, the system receives approximately 32 thousand accesses per day, which shows a certain robustness of the implemented solution.

The data collected show that most visitors accessed the site through searches in *Google*. On most occasions, the items sought by these visitors are names of people, companies or their official identifying numbers. Since these are items which are hardly ever found outside an official gazette, it meant that DL is among the few results that are returned from such searches.

Although PDFs are apparently subject to indexation by search engines, the official site of the publication of the DOM does not seem to be indexed by them. Bearing in mind that this site does not have a *robots.txt* (the file that can prohibit such indexation), it is believed that this did not occur, since there are no direct links to the PDFs. They can only be visualized by using the search interface on the site. In DL, the search filters are links that lead to the pages with the results,

which in turn, lead to the published articles. This difference seems to have been essential to allow the indexation of the content.

5.2 Users Feedback

Regarding the online questionnaire on the DL website, 105 complete answers were obtained. From the 7 users who experienced difficulty with DL (about 7 % of the total), 4 stated that they had not been able to locate some of the features on the interface, 2 had problems with the security certificate and 1 did not specify the reason. From these 7, only 2 (2 %) stated that they had not had problems with the official IO website. In contrast, 39 users (37 %) had difficulties with the IO website and 33 of these (31 % of the total) did not have any difficulty with the DL. The problems found in the IO website included: seeking terms within the PDFs; copying the text from the PDFs; printing only part of the texts; the fact that the small font makes it difficult to read.

Finally, 65 respondents (62 %) were satisfied with the DL, as opposed to 13 (12 %) who were dissatisfied. These data support the hypothesis that DL is easier to use than the IO website.

Beside this, it is worth mentioning a few reports that provide some insight about the impact of Diário Livre on three distinct groups: civil servants, social movements and hacktivists. The first report illustrates the potential saving of resources (both financial and human) provided by using the Official Gazette in an open format. One civil servant from the legal area took approximately 30 min daily to read and select the subjects in the paper version of the Official Gazette. Using the DL API through a data-reading script, it was possible to automate the work and, in a few seconds, a report is produced and forwarded by e-mail to the interested parties.

Moving beyond the internal use of the City Hall, there are reports of usage by the organized civil society: a president of the neighborhood association in São Paulo stated he used Diário Livre to clarify his doubts and help him with the requests the neighbors in his county send to the association. He mentioned examples of complaints related to stores, since it is possible to easily seek information on operating licenses and their situations in the open version of the Official Gazette.

It is also important to report that there are examples of initiatives that are seeking to integrate applications with the data in Diário Livre. A group of local civil hackers integrated a tool on the platform for their project on budget transparency. This type of initiative shows there is a potential to stimulate the creative economy of applications from the development and improvement of Diário Livre.

6 Conclusion

This article described the joint initiative between the public power and academia to offer the Official Gazette of the city of São Paulo in a digital format that

followed the principles of open data and thus foster the reuse of information by a broader range of possibilities.

The project was successful in reaching its primary objective, managing to make data available in real time and in different open formats, reaching the *4 stars of open data*. Although not yet an official portal, the initiative has become an important proof of concept that has not only demonstrated the technical viability of making an official gazette available in open format, but also offered a tangible example of the contribution of open format to a broader public.

The methodology employed for action research involved holding several meetings with the team and (together with the first event and related bibliography) made it possible to conduct a survey of the required conditions for the data publishing tool. This was then projected, implemented and launched with the same partnerships during the second event.

It was confirmed that the process fulfilled several of the recommendations made in the literature about open government data. This particularly applied to the tool produced and although it failed to satisfy all the possible uses for the data, it can be said that it met the essential requirements. Moreover, it was rated highly by the users and public administrations, as well as obtaining a growing number of accesses.

With regard to the factors in the initiatives that have motivated and attracted the key players, one can cite the search for greater transparency and compliance with related legislation. The challenge that arose largely originated from the fact that open government data initiatives at a local level are relatively recent and require a number of technological and cultural adjustments.

From the stand point of participation, an attempt was made to take this into account at every stage of the process. Despite the constraints imposed, which perhaps led to a lower level of participation than had originally been desired, it can be considered that in this respect, the final result was satisfactory. Two events were held, the first of which assisted in meeting the requirements of the final tool and the second helped to obtain a general equilibrium in the process. In addition, the online questionnaire also made a final means of contact possible and this was used by more than a hundred users.

The system had some repercussions in the media and received recognition from specialists in innovatory public management in the form of a reward at an official public event. The publication of the data also allowed some attempts to be made for their reuse both by the public administration and by people outside. These outcomes have already shown the potential of the database for obtaining non-systematized information at any place, although they have also revealed the challenges facing the question of the reuse of these data.

The implementation of the proof of concept and its consequent repercussion was a first step, and further works point towards directing the publication as linked data, with the crossing with other government database, such as, for instance, the one regarding public contracts and purchases.

Although the data contained in the Official Gazette deal with subjects of great diversity, which hinders the creation of a common vocabulary or an

ontology to represent its content, it is possible to create ontologies or data connected to certain aspects dealt in the gazette. The creation of a system based on "Linked-Data" is being created to ease the recovery of information related to tenders, and the authors expect to obtain positive results both for public management and for citizenship in a broader context.

Finally, there is another question regarding the effects that these data can exert either on the public administration itself or on the diverse groups that comprise society – hackers, journalists, academics,OSCs, social movements, people with little familiarity with technology and others. To achieve this, we intend to conduct an analysis of the profile of the users with the aim of improving the presentation and attempting to design an adaptive version of our interface.

References

1. Benkler, Y.: The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press, London (2006)
2. BRASIL: Decreto nº 4.520. 16 de dezembro de 2002. http://www.planalto.gov.br/ccivil_03/decreto/2002/D4520.htm
3. Sempla: Diário Oficial da Cidade de São Paulo: manual de instruções (2010). http://www.prefeitura.sp.gov.br/cidade/secretarias/upload/chamadas/manual_de_instrucoes_do_diario_oficial_2010_1306171567.pdf
4. THE 8 PRINCIPLES OF OPEN GOVERNMENT DATA (2007). http://opengovdata.org
5. Berners-Lee, T.: Is Your Linked Open Data 5 Star? (2010). http://www.w3.org/DesignIssues/LinkedData.html
6. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. **29**(4), 258–268 (2012). http://www.tandfonline.com/doi/abs/10.1080/10580530.2012.716740. Accessed 11 Nov 2013
7. Open Knowledge. http://opendefinition.org/od
8. Lakomaa, E., Kallberg, J.: Open data as a foundation for innovation: the enabling effect of free public sector information for entrepreneurs. IEEE Access **1**, 558–563 (2013)
9. Ubaldi, B.: Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives (2013)
10. Halonen, A.: Being Open About Data: Analysis of the UK Open Data Policies and Applicability of Open Data. The Finnish Institute in London, London (2012)
11. Davies, T., Perini, F., Alonso, J.: Researching the Emerging Impacts of Open Data, vol. 20. World Wide Web Foundation, Washington, DC (2013)
12. Gurstein, M.: Open data: empowering the empowered or effective data use for everyone? First Monday, **16**(2) (2011). http://firstmonday.org/ojs/index.php/fm/article/view/3316/2764
13. Benjamin, S., Bhuvaneswari, R., Rajan, P.: Bhoomi: 'E–governance', or, an anti-politics machine necessary to globalize Bangalore? In: CASUM-m Working Paper (2007)
14. Grimmelikhuijsen, S.: A good man but a bad wizard. About the limits and future of transparency of democratic governments. Inf. Polit. Int. J. Gov. Democr. Inf. Age **17**(3/4), 293–302 (2012). http://search.ebscohost.com/login.aspx?direct=true/&db=afh/&AN=84341711/&lang=pt-br/&site=ehost-live

15. Richards, R.C. (2010). http://legalinformatics.wordpress.com/2010/07/26/federal-register-2-0-now-available
16. Cifuentes-Silva, F., Sifaqui, C., Labra-Gayo, J.E.: Towards an architecture and adoption process for linked data technologies in open government contexts: a case study for the library of congress of chile. In: Proceedings of 7th International Conference on Semantic Systems, pp. 79–86 (2011) http://doi.acm.org/10.1145/2063518.2063529
17. Brandao, S.N., Rodrigues, S.A., Silva, T., Araujo, L., Souza, J.: Open government knowledge base. In: ICDS 2013, 7th International Conference on Digital Society, pp. 13–19 (2013). http://www.thinkmind.org/index.php?view=article&articleid=icds_2013_1_30_10168
18. Davies, T.G.: Open data policies and practice: an international comparison (2014). SSRN 2492520, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2492520
19. Linders, D.: From e-government to we-government: defining a typology for citizen coproduction in the age of social media. Gov. Inf. Q. **29**(4), 446–454 (2012)
20. Dudley, L.R.: Federal Register 2.0: public participation in the Twenty-First Century. Legis. Pol'y Brief **3**, vii (2011)
21. Baskerville, R.L.: Investigating information systems with action research. Commun. AIS **2**(3es), 4 (1999)
22. Kim, P.H.: Action research approach on mobile learning design for the underserved. Educ. Technol. Res. Dev. **57**(3), 415–435 (2009)
23. Reid, C.: Advancing women's social justice agendas: a feminist action research framework. Int. J. Qual. Methods **3**(3), 1–15 (2004)
24. IMPRENSA OFICIAL: Manual de Conversão para PDF Envio de Arquivos ao Diário Oficial (2011). https://pubnet.imprensaoficial.com.br/pubnetii/manuais/ManualGeracaoPDF.pdf
25. Smiley, D., Pugh, E.: Solr 1.4 Enterprise Search Server. Packt Publishing Ltd., Birmingham (2009)
26. Grainger, T., Potter, T., Seeley, Y.: Solr in Action. Manning, Greenwich (2014)
27. DuPlain, R., Balser, D.S., Radziwill, N.M.: Build great web search applications quickly with Solr and Blacklight. In: Proceedings of SPIE, vol. 7740, pp. 774011–774011-12 (2010). http://dx.doi.org/10.1117/12.857899