A Collaborative Approach to Support Document Structuring Process in the Context of Open Government Data

Andreiwid Correa

Introduction

1 Public organizations have found in Open Government Data (OGD) a way to enhance accountability, to support decision-making activities and to lead to innovative applications constantly demanded from society. OGD is a government-oriented variant of Open Data, which in turn is a major component of what is called the Data Revolution that basically means the unlocking of enormous volumes of information to be freely used, reused and redistributed by anyone (Tauberer, 2014). Some argue that Open Data is increasingly recognized as a new form of infrastructure that is transforming how governments, businesses, and citizens are organized in an increasingly networked society (Open Data for Development Network: Building an inclusive data revolution, 2015).

2 Data openness occurs when data is published in a way that complies with (among others) *accessible* and *machine processable* open data principles (OpenGovData, 2007; Tauberer, 2014). It means that data publishing with the use of unstructured documents and non-open formats hampers data consumers as they lack openness. A study estimated that roughly 13% of published files in some main open data portals around the world have their data made available in PDF (Andreiwid Sheffer Corrêa & Zander, 2017). A recent assessment of data openness in Brazilian municipalities discovered that HTML is the most widely preferred format followed by PDF, where both formats represented 79% of all published documents (Andreiwid Sh. Corrêa, Paula, Corrêa, & Silva, 2017).

3 Apart from the apparent benefits of publishing data through PDF and HTML, these formats lack essentially machine-readability feature and not everyone knows how to

extract data from them. Thus, the data processing and acquisition become costlier and challenging as there is no sufficient information regarding structure of the data contained in it (Silva, 2010).

4 In a previous work (Andreiwid Sheffer Corrêa, Corrêa, & Silva, 2015), we have introduced a conceptual layered software architecture that envisaged a collaborative structuring of information into open data. This paper aims at expanding one of this architecture's layer by elaborating the collaborative approach and emphasizing its business model.

5 The contribution of this work is seen in the form of software requirements whose design was inspired by Wikipedia project adapted to the context of open government data. These requirements would guide the implementation of fully operational software that allows the participation of users from open data community while government agencies prepare themselves to publish open data by default.

Related work

6 In a previous work (Andreiwid Sheffer Corrêa, Corrêa, & Silva, 2015), we have introduced a conceptual four-layered software architecture that envisaged structuring of information into open data. One of the architecture's layers is called "Collaborative" and it was supposed to allow users interact with unstructured or non-open data sources by providing tools and proper user interface. But the architecture specification did not provide enough information on how it would handle the participatory process and what mechanism would be developed to engage users to collaborate.

7 In this way, we investigated something similar and widespread that is the Wikipedia open content encyclopedia. Articles in Wikipedia can be proposed and improved by any user from their knowledge in a certain area. Since its launch in 2001, Wikipedia has been among the most ingenious collaborative projects online and has millions of articles in hundreds of languages.

8 To understand the participatory process that influenced Wikipedia, Bryant, Forte & Bruckman (2005) conducted a multi-user survey that contributed significantly to the understanding of this growing community. The authors identified several findings of which some are useful to understand, construct and advance in the elaboration of the collaborative approach envisaged by our software architecture:

- The users who edited an article for the first time did so on matters of their personal interest. Initially, the objective was to correct small problems such as omissions and inaccuracies in the text. At this stage, users viewed themselves as consumers rather than contributors to the community.
- In their early experiences, users did not have the real sense of collaborative community. Interaction with other users was also scarce. As they increased their involvement, they learned more about community rules and also interacted with other contributors.
- Over time, users evolved in the community by increasing their level of involvement. They started from the so-called peripheral zone of contribution to the central part of the community, where activities and interactions became more complex because they involved discussions and decisions that aided the growth and quality of the contributions network, such as avoiding vandalism in controversial articles and highlighting articles in Featured.
- The users contributed because they believed in what the community produced. The authors noted that the motivation to contribute could come from other sources such as the

expectation of community reciprocity in the future and sense of reputation, which did not depend on altruism pure and simple.

- Users reported that the first thing they did when they entered the tool was to check the watch list to check for changes in their articles of interest in order to quickly identify vandalism.

9 Inspired by these observations, in the next section we describe the collaborative approach in the form of software requirements necessary to guide the implementation of fully operational software that allows the participatory process in the context of open government data.

Design and methodology

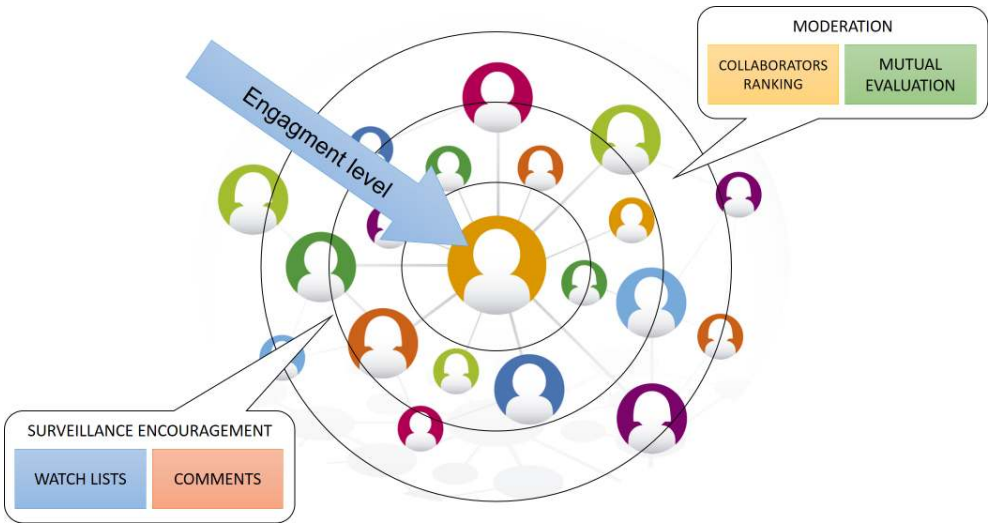
10 The main reason of a network of collaborators is to handle the numberless of unstructured and non-open data sources existing in the different levels of government. At this point, government agencies are still in charge of disclosing information in detriment of open data, mainly because of technical limitation of their staff or of misunderstanding of data openness and its principles.

11 In this context, we introduce data structuring as the main process of turning information into open data. Instead of documents with unstructured and non-open data sources, there are datasets available for consuming by anyone for any purpose. Thus, collaborators conduct the structuring process in order to make them accessible and machine processable.

12 Collaborators can be encouraged by the dissemination of content in the press that demands opening data, *e.g.* when an agency discloses its budgetary information. Thus, a network of professionals such as journalists, political analysts, economists would be in charge of structuring data based on their specific interest, besides the possibility to ordinary citizens contribute as well.

13 The proposed collaborative approach takes into consideration two main requirements of this network that catalyzes the engagement level of collaborators. They are moderation and surveillance encouragement, both illustrated in the following figure.

Overview of the collaborative approach and its main requirements



(Elaborate by authors)

14 The moderation is build on two main requirements which implement collaborators ranking and mutual evaluation. In this way, the software expects from user's initial contribution with data in their area of interest *e.g.* by structuring data from their city until becoming recognized experts and playing a central role in the network. Also, collaborators can evaluate the structuring performed by other users which demands software functionalities to handle the mutual evaluation and then feed the collaborators ranking. In this scenario, as the number of collaborators grows, the level of engagement tends to increase as well, consequently quantity and quality of contributions become significant.

15 Surveillance encouragement supports the contributions from users by implementing watch lists and comments to the community. In this way, software needs to handle alerts that fire when a certain collaboration appears on the system. Comments exist in the form of short texts where users discuss to each other the validity of a contribution that will be used by the network.

16 With the implementation of these requirements, the design of a software with the collaborative approach should provide the following to the open data community:

- Reusable data–characterized by the availability of datasets that can be downloaded, remixed or crossed with other data in order to verify consistency with other data sources;
- Machine-processable data–characterized by the availability of open and interchangeable datasets for processing by any tool of interest to users;
- Access APIs–characterized by the availability of automated resources to allow access and extension of the features by the development of other functionalities based on those currently offered.

17 We report herein this in progress work is under development of a software prototype that implements the proposed collaborative approach. It was developed using a web system as backend and an add-on integrated with one of the most used internet browser Mozilla Firefox. The backend essentially provides APIs and database systems to support functionalities. The add-on serves as user interface where collaborators conduct the data structuring process.

18 In the next action we are considering the involvement of open data community to help in the development and dissemination of the prototype. Analysts and software developers are welcome to join in and increase the number of resources to make open data part of our lives.

Expected outcomes

19 This paper aimed to expand the collaborative approach taken into consideration in a previous work regarding a software architecture designed to overcome the current obstacles in the availability of open government data operated by most of the current government agencies.

20 The expected outcomes involved in the architecture and its underlying collaborative approach meet the need of public interested in consuming data, especially from open government data community that militates for the opening of public records. With the architecture, and its collaborative approach involved, it will be possible to guide the implementation of fully operational software systems that will provide users with tools to easily open data from any agency.

21 Thus, the aim of building a network of collaborators is to help in the structuring process, as the numberless of unstructured documents and non-open formats make it difficulty for citizens to consume data. The expected outcomes in this matter show as a temporary solution to push governments to the data openness.

BIBLIOGRAPHY

References

Bryant, S. L., Forte, A., & Bruckman, A. (2005). "Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia." In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, 1–10. New York, NY, USA: ACM. https://doi.org/10.1145/1099203.1099205

Corrêa, Andreiwid Sh., Paula, E. C. de, Corrêa, P.L.P., & Silva, F. S. C. da. (2017). "Transparency and open government data: a wide national assessment of data openness in Brazilian local governments." *Transforming Government: People, Process and Policy*, *11* (1). https://doi.org/10.1108/TG-12-2015-0052

Corrêa, Andreiwid Sh., Corrêa, P.L.P., & Silva, F.S. C. da. (2015). "A Collaborative-oriented Middleware for Structuring Information to Open Government Data." In *Proceedings of the 16th Annual International Conference on Digital Government Research*, 43–50. New York, NY, USA: ACM. https://doi.org/10.1145/2757401.2757409

Corrêa, Andreiwid Sh., & Zander, P.-O. (2017). "Unleashing Tabular Content to Open Data: A Survey on PDF Table Extraction Methods and Tools." In *Proceedings of the 18th Annual International*

Conference on Digital Government Research, 54–63. New York, NY, USA: ACM. https://doi.org/10.1145/3085228.3085278

Open Data for Development Network: Building an inclusive data revolution. (2015). *Open Data for Development*. Retrieved from http://od4d.net/wp-content/uploads/2016/06/OD4D_annual_report_2015.pdf

OpenGovData. (2007). *Eight Principles of Open Government Data-OpenGovData.org*. Retrieved November 27, 2014, from http://opengovdata.org/

Silva, A.C. (2010). *Parts that add up to a whole: a framework for the analysis of tables* (Ph.D.'s Thesis). Edinburgh University, UK.

Tauberer, J. (2014). *Open Government Data: The Book–Second Edition.* Retrieved from https://opengovdata.io/

ABSTRACTS

The online availability of public data using unstructured documents and non-open file formats is still found in many government agencies what hampers society to consume data, as unlocking data from them is not a trivial task directed to everyone. This in progress work aims at expanding a previous proposal by elaborating an important aspect of a conceptual software architecture which is the collaborative approach in the context of open government data. The contribution of this work is shown in the form of software requirements that make it possible for users in the open data community be involved with the entire data structuring process while government agencies prepare themselves to publish open data by default.

INDEX

Keywords: Open Data, PDF, HTML, Software Architecture, Collaborative

AUTHOR

ANDREIWID CORREA

Federal Institute of Sao Paulo, Brazil
andreiwid@gmail.com
(corresponding author)