EPUB 2018

The 22nd International Conference on Electronic Publishing

Connecting the Knowledge Commons:
From Projects to Sustainable Infrastructure

June 22-24, 2018 | University of Toronto, Canada

A collaborative approach to support document structuring process in the context of open government data

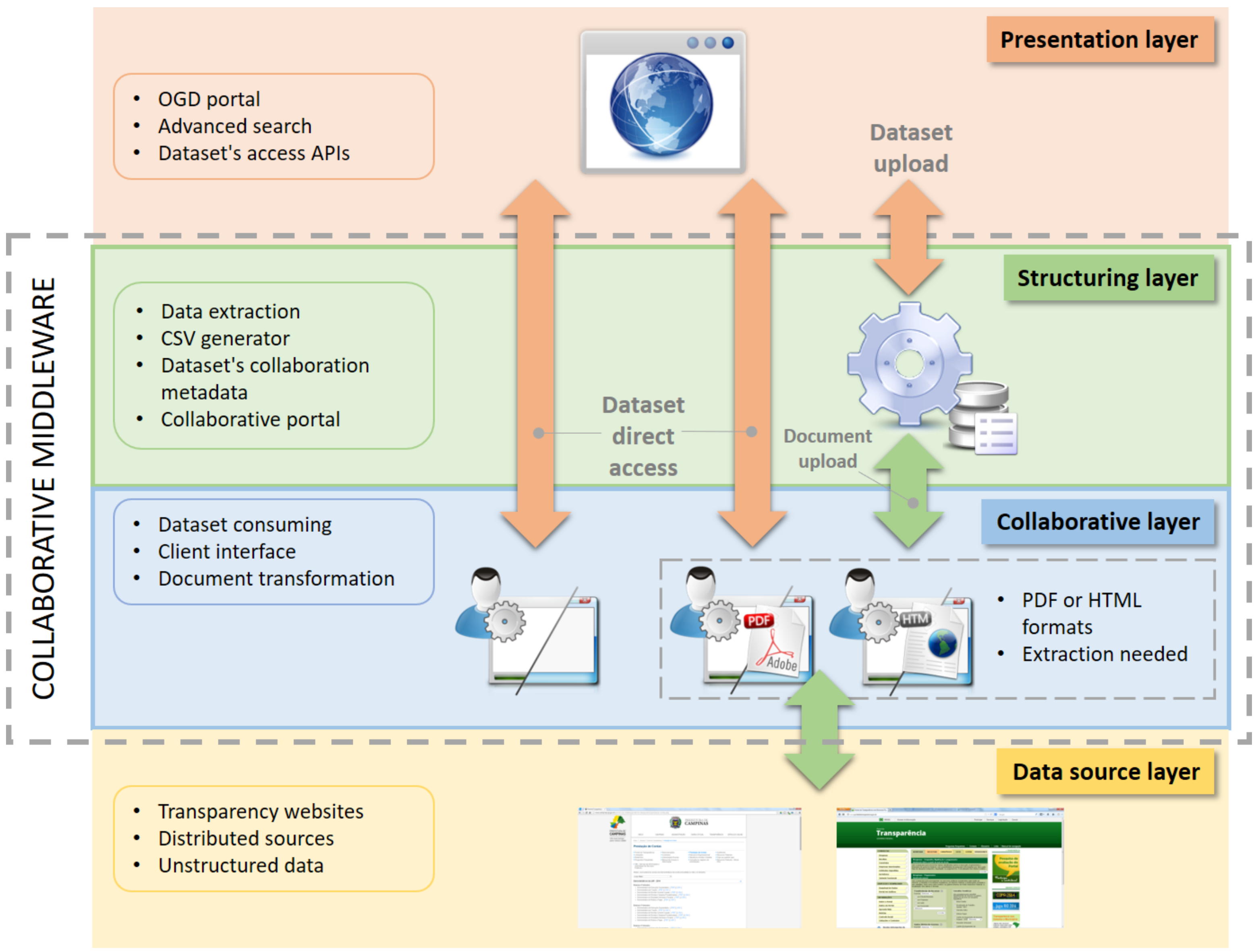
Andreiwid Sheffer Correa - *Federal Institute of Sao Paulo*
andreiwid@ifsp.edu.br

ABSTRACT

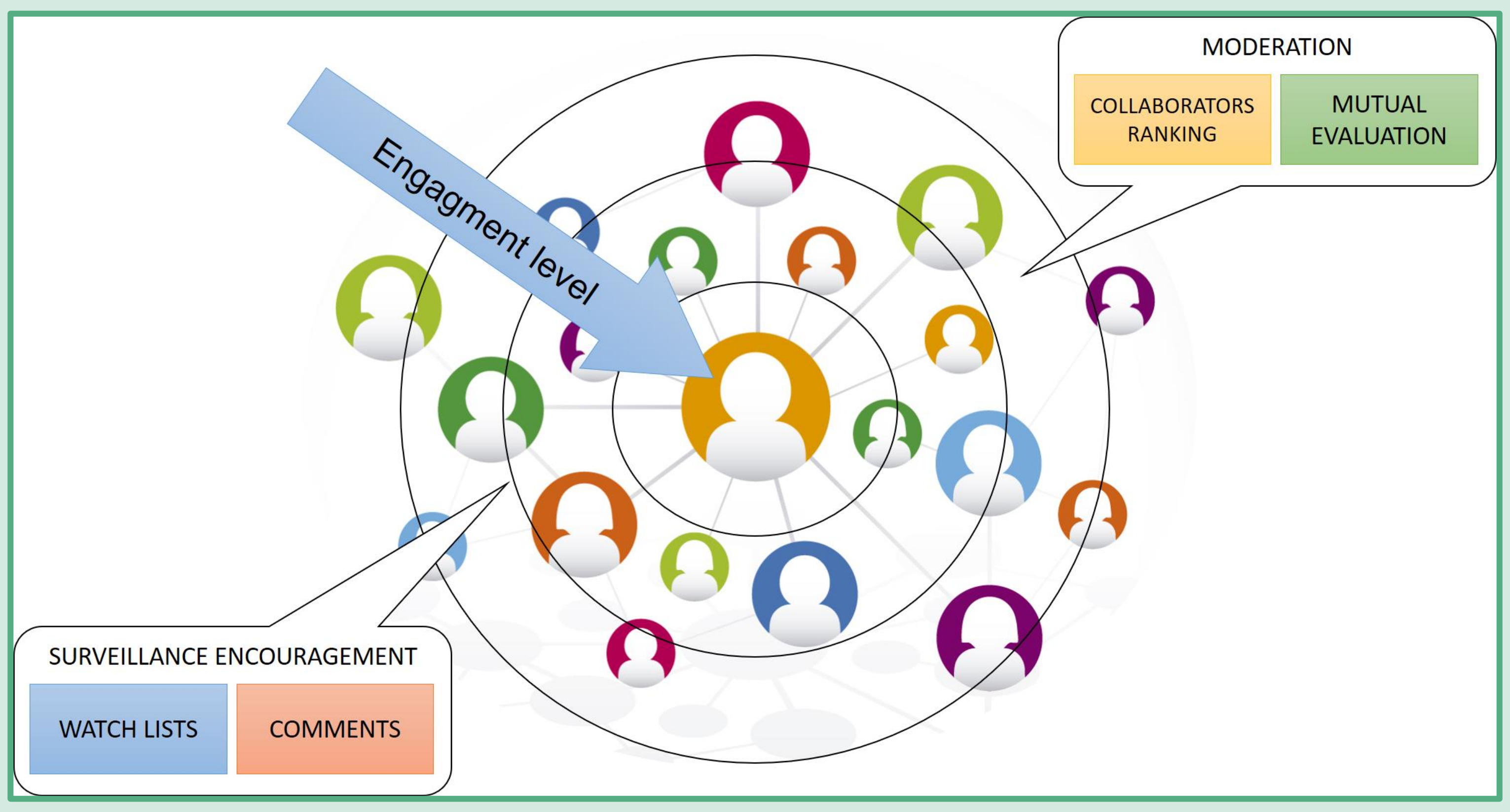
- Unstructured documents (*e.g.* PDF) and non-open file formats (*e.g.* XLS) are still found in the context of Open Government Data
- As they do not comply with Open Data, unlocking data from them is not a trivial task directed to everyone
- It was defined a software architecture that envisaged structuring of information into open data
- This work elaborates the collaborative approach to engage the open data community in the structuring process
- Contributions of this work are shown in the form of software requirements
- Expected outcomes include elements to guide implementation of fully operational software systems that provide to users tools to easily open data from any data source

BACKGROUND

- Data openness occurs when data is published in a way that complies with (among others) *accessible* and *machine processable* open data principles
- It is estimated that roughly 13% of published files in some main open data portals around the world have their data made available in PDF [1]
- Dynamic-generated HTML is the most widely preferred format followed by PDF, where both formats represented 79% of all published documents in Brazilian cities [2]
- PDF and HTML lack essentially machine-readability feature
- It was introduced a conceptual **layered software architecture** that envisaged a collaborative structuring of information into open data [3]



DESIGN & METHODOLOGY



- The main reason of a network of collaborators is to handle the numberless of unstructured and non-open data sources at any levels of government
- Collaborators conduct the structuring process in order to make data accessible and machine processable
- They can be encouraged by the dissemination of content in the press that demands opening data, *e.g.* when an agency discloses budgetary information
- The proposed collaborative approach is based on two main requirements: **moderation** and **surveillance encouragement** [4]

EXPECTED OUTCOMES

- Involve the open data community in the structuring process
- Support process and provide tools to foster greater civic participation thus meet the need of public interested in consuming data
- Guide the implementation of fully operational software systems to easily open data from any agency
- Define a model as well as a solution (even temporarily) to push organizations to the data revolution by opening public records

REFERENCES

[1] Corrêa, Andreiwid Sheffer, & 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 (pp. 54–63). New York, NY, USA: ACM. https://doi.org/10.1145/3085228.3085278

[2] 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

[3] Corrêa, Andreiwid Sheffer, 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 (pp. 43–50). New York, NY, USA: ACM. https://doi.org/10.1145/2757401.2757409

[4] 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 (pp. 1–10). New York, NY, USA: ACM. https://doi.org/10.1145/1099203.1099205

This work is supported by

INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico