Hippolyta

a framework to enhance open data interpretability and empower citizens

Barcellos, Raissa; Bernardini, Flavia; Viterbo, Jose; Zuiderwijk, Anneke

DOI
10.1145/3598469.3598559

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 24th Annual International Conference on Digital Government Research - Together in the Unstable World

Hippolyta: a framework to enhance open data interpretability and empower citizens

Raissa Barcellos
Fluminense Federal University
Niterói, RJ, Brazil
raissabarcellos@id.uff.br

Flavia Bernardini
Fluminense Federal University
Niterói, RJ, Brazil
fcbernadini@ic.uff.br

José Viterbo
Fluminense Federal University
Niterói, RJ, Brazil
viterbo@ic.uff.br

Anneke Zuiderwijk
Delft University of Technology
Delft, South Holland, Netherlands
a.m.g.zuiderwijk-vaneijk@tudelft.nl

ABSTRACT

Open government data initiatives have been rising quickly in recent times. They are encouraged by a wish to democratize data access and knowledge production and enhance cities socially and economically. The hardship of interpreting data can be considered an obstacle to using open government data and more prominent citizen engagement. Technology is crucial to enhance data interpretability and the practical construction of an open government. Nevertheless, the literature needed an instrument to support open government data's interpretability. In this work, our primary goal is to present the definition, implementation, and evaluation of a framework named Hippolyta, which is qualified to help citizens to interpret open government data. Hippolyta first identifies the citizen's necessities using a semantic enrichment module. After this step, the framework conducts the data collection through the same data retrieval module. Finally, Hippolyta creates a graphic visualization through a data visualization module. This study is relevant since it furnishes comprehensive insights into what the open data interpretability concept is composed of and which framework modules can sustain open data interpretation.

CCS CONCEPTS

• **Applied computing** → **E-government**.

KEYWORDS

Data Interpretability, e-government, open data

ACM Reference Format:
Raissa Barcellos, Flavia Bernardini, José Viterbo, and Anneke Zuiderwijk. 2023. Hippolyta: a framework to enhance open data interpretability and empower citizens. In *24th Annual International Conference on Digital Government Research - Together in the unstable world: Digital government and solidarity (DGO 2023), July 11–14, 2023, Gdańsk, Poland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3598469.3598559

DGO 2023, July 11–14, 2023, Gdańsk, Poland

ACM ISBN 979-8-4007-0837-4/23/07…$15.00
https://doi.org/10.1145/3598469.3598559

1 INTRODUCTION

The global open government data movement has evolved significantly in the last decade, and it has been embraced by governments worldwide with varying degrees of enthusiasm [22]. Government is a source of very particular types of data. While the future of open data may be an expanding and changing one, at its core will remain the importance of governments as a source of quality, accessible, and reusable data that can drive objectives of transparency and accountability, stimulate innovation and increase citizen engagement [22]. However, open government data still need to enhance citizen interaction and engagement [29]. Open data portals often maintain valuable data with the potential to impact the citizen directly. The correct interpretation of the available datasets is crucial in empowering all citizens [29].

The difficulty of interpreting data can be considered a social barrier to using open government data and greater citizen engagement [30]. Moreover, the capacity to interpret open data is essential to exploring and understanding the actual value of data [14]. However, in previous works, we found that OGDPs still need to support the user in interpreting the available data, even an audience with extensive data science skills [4]. Also, we point out that currently, OGDPs need to implement more computational tools to efficiently support citizens in interpreting the available data [4]. Computational tools are essential to enhance data interpretability and the practical construction of an open government [4].

In this work, our main objective is to propose the definition, implementation, and evaluation of a framework named Hippolyta, which is capable of helping citizens to interpret open government data. We can define data interpretability as the capability of an accurate, complete, consistent, coherent, and organized dataset to convey significance to the user, stimulating the formation of his knowledge and engagement, from a simple and clear language [4] . Applying strategies and tools to identify, manage and disseminate data becomes fundamental for data interpretation [15]. Organizations or portals that provide datasets with low representational quality — such as lack of metadata and inconsistent formats — have more incredible difficulty in facing the interpretation deficit [14]. Some actions help to reduce the interpretation deficit, such as: (i) standardizing data, using metadata and accessible formats; (ii) using semantic analysis and sentiment analysis techniques, and (iii) using data visualization techniques in order to facilitate the agile understanding of insights, such as trends and relationships between

datasets [14]. Furthermore, we can consider Hippolyta as a suggestion for changing some steps of the standard data manipulation process.

This study is societally relevant since open government data can potentially lead to the generation of societal, economic, and operational value (e.g., see [3] and [17]). The difficulty to interpret open government data well is an essential barrier in the process of value creation with open government data. This study addresses the interpretability barrier and its outcome provides insights that can potentially make open data use easier for citizens. Moreover, this study is scientifically relevant since it provides in-depth insights into what the open data interpretability concept is composed of and which framework modules can support open data interpretation. By doing this, it contributes to the few studies on open data interpretability research [e.g., [4] and unravels the dimensions of open data interpretability.

This work is organized as follows:

2 BACKGROUND

2.1 Data Interpretability

Citizens need to be able to correctly interpret the available data to participate more actively in democratic processes [24]. OGD programs aim to contribute to public transparency. Consequently, governments openly share their data with the public so that citizens can hold the government accountable and better understand how the government acts. In order to attain the objectives of transparency and accountability, governments generally assume that citizens can interpret their data. However, there still needed to be a formal definition in the literature about data interpretability. In this previous work [4] we formally conceptualize data interpretability as the capability of an accurate, complete, consistent, coherent, and organized dataset to convey significance to the user, stimulating the formation of his knowledge and his engagement, from a simple and clear language [4]. Also, we grouped some characteristics that help us to constitute the concept of data interpretability, considering the context of OGDPs [4].

G1 **Understandability, simplicity, clarity and readability** : This set of characteristics is related to the method of data presentation, necessary to maintain the user's interest in consuming the data.

G2 **Reliability and traceability**: This set of characteristics is related to the possible disbelief of users about the data presented, being a minimum condition for use but not essential to the interpretation.

G3 **Structuring, organization**: This set of characteristics is related to the fact that poorly organized datasets make the task of information retrieval difficult, making obtaining value from these data more complex.

G4 **Accuracy, correctness**: This set of characteristics is related to the fact of obtaining a misinterpretation, distorted. However, according to [28], knowledge requires trust, so for us to reach a valid interpretation, the dataset must remain correct, unbiased, and accurate.

G5 **Completeness**: Completeness is related to the importance of detailing the dataset.

G6 **Conciseness**: Conciseness is related to the proper way to remove unnecessary elements from a dataset — such as avoiding similar naming data differently — which can become a problem if it affects the completeness characteristic.

G7 **Consistency, coherence**: This set of characteristics refers to the condition of the data's ability to fulfill, without contradiction, all the properties of integrity, equivalence, logic, authenticity, and standardization.

G8 **Informativity**: Informativeness is related to maintaining the user's interest and satisfaction.

In order to complement the data interpretability definition, we also built a model for leveraging data interpretability in OGPDs, as shown in Figure 1. This model can provide a basis for leveraging interpretability in OGD. In the interpretability model for open government data, OGDPs must perform actions such as: uncomplicate the data, track the data, organize the data, adjust the data, complete the data, synthesize the data, adapt the data, and inform citizens. After the publication of the formal definition, other works in the area are already using the data interpretability concept [10, 20, 25].

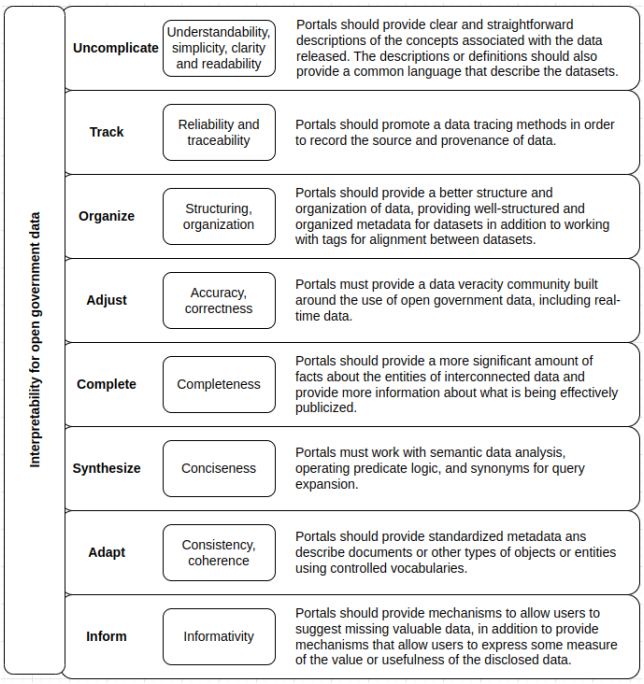

Figure 1: Model for leveraging data interpretability in OGPDs [4]

2.2 Citizen-Sourcing

Citizen-sourcing is the adoption of crowdsourcing principles and technologies in the public sector, functioning as an instrument to support citizen participation and knowledge sharing [23]. Recently, governments worldwide are pursuing strategies to expand citizen participation and collaboration [26]. In the context of *Citizen-sourcing*, the data collection process starts with an enrollment request in the form of an open call. Several actors may be able to

make a request, such as (i) government representatives, (ii) non-governmental organizations, (iii) or citizens themselves. This request is then published on an online platform that acts as an intermediary between requesters and providers. Citizens can also act as testers and rate submitted apps by reviewing them, posting possible improvements for developers, and providing a rating among other contributions [26]. Citizen-sourcing can provide more available services to society, to offer services where the citizen is at the center of the data flow [23].

Citizen-sourcing has its roots in three distinct phenomena: (i) the new technologies of Web 2.0, which allow citizens to communicate with governments with minimal transaction costs; (ii) the success of decentralized and distributed development of products and content in the private sector through open source, open innovation and crowdsourcing and (iii) the "top-down" approach of former US president, who facilitated open government by prioritizing participation and collaboration with citizens [1].

In this work, we will use the concept of citizen-sourcing when we explore the results of the Hippolyta evaluation, specifically in Section XX.

3 RELATED WORKS

Our primary motivation for building this related works research is to address that OGDPs must offer computational resources to generate a more added value of excellence to the citizen based on the available data. In this sense, efficient architectures, frameworks, processes, and methodologies are necessary to identify, map, develop and plan resources for improved interaction between citizens and OGDPs. Some initiatives addressed in the literature contribute to filling this gap. We found works that promote more efficient interfaces, providing differentiated functionalities that facilitate understanding the data by different user profiles and provide data visualization options.

Cantador *et al.*, in [5], present a chatbot to access open government data. The chatbot developed allows searching and exploring datasets. Exploration is done through complex queries that non-expert users quickly construct through natural language conversation. First, the user accesses and interacts with the chatbot through conversations in an instant messaging application. Soon after, a Natural Language Processing component extracts entities and identifies the target intent. If the intent is to search a collection, the chatbot prompts the user for terms associated with relevant datasets. It uses the provided terms to initiate a keyword-based search against a database index. As a result of the process, the chatbot presents a list of datasets on the input keywords. The chatbot maintains a conversation with the user and asks the user for the elements needed to create an SQL query that represents his information needs. Once the chatbot creates the query, it is posted to the database, retrieving the data items of interest. And then, in the end, these items are finally presented to the user. In addition, the authors report a study carried out to evaluate chatbots according to the achievement of a series of public service values, in addition to measuring different objective and subjective metrics. Experimental results show that the proposed system outperforms traditional methods followed in OGDPs.

The authors Chokki *et al.*, in the work [6], identify a list of features needed in the design of a generic tool for data storytelling and then implement these features into a usable tool called ODE — Open Data Explorer. ODE provides additional resources to facilitate data storytelling by users, such as (i) direct connection to portals, (ii) estimation of data quality, (iii)) data overview, (iv) recommended viewing of selected data, and (v) feedback collection. The authors also conducted interviews with eleven users to evaluate whether the ODE is easy to use and valuable at all stages of data storytelling, but also to collect suggestions for additional features to be implemented. Unlike generic data visualization tools, ODE provides users with an end-to-end tool to transform data into information without using separate tools. The ODE also allows users to give their feedback on data visualizations and later use it to improve the rules of initial data visualizations.

Arribas-Bel*et al.* [2] develop the notion of "open data product" and define an open data product as "the open result of the processes through which a variety of data are turned into accessible information through a service, infrastructure, analytics or a combination of all of them, where each step of development is designed to promote open principles". So, they contribute to the open data literature by providing a framework that expands the notion of how open data can be generated and what can constitute the basis to generate open datasets, as well as how to ensure its final usability and reliability.

4 RESEARCH DESIGN

To reach our goal, we (i) deepen concepts such as data interpretability and citizen-sourcing, (ii) carried out a literature review in order to highlight works that address architectures, structures, processes, and methodologies to identify, map, develop and plan resources for improving the interaction between citizens and open government data portals (OGDPs), (iii) we defined the architecture of the framework, named Hippolyta, (iii) we evaluated the Hippolyta instantiation, in order to consolidate the efficiency of the framework in terms of improving the interpretability of open government data. Moreover, we can consider the framework as a suggestion for changing some steps of the standard data manipulation process.

Hippolyta first identifies the citizen's needs using a semantic enrichment module. After this step, the framework performs the data collection through the same data retrieval module. Finally, data visualizations are performed through a data visualization module. Hippolyta's code is available at [4]

In previous works [4], we already evaluated the semantic enrichment and data retrieval modules. However, we still need to evaluate whether Hippolyta can promote an improvement in the interpretability of open government data. We chose to use the focus group technique to evaluate Hippolyta. The focus group is a qualitative research method that works as a type of in-depth interview carried out in a group [19]. The focus or object of analysis is the interaction within the group [19]. In the focus group, participants influence each other through their responses to ideas and contributions during the discussion. The moderator encourages discussion with comments or subjects. The fundamental data produced by this technique are the transcripts of the group discussions and moderator reflections and notes. The general characteristics of a focus group are the involvement of people, a series of meetings,

the homogeneity of participants concerning research interests, the generation of qualitative data, and discussion focused on a topic, which is determined by the objective of the research [19].

The advantages of using the focus group to evaluate Hippolyta are: (i) easy conduction, allowing better exploration of interpretability factors and generation of hypotheses; (ii) opportunity to collect group interaction data, which focuses on our topic of interest; (iii) it has a low cost compared to other methods; (iv) greater agility in providing results; (v) the methodology is flexible and has high face validity, which means that we measure more easily what we intend to measure [11].

The focus group for evaluating Hippolyta had five members, except for the moderator, Table 1 summarizes information about the members. The group meetings were divided into two sessions. The script was organized so that all characteristics, that define the concept of interpretability, were discussed.

Table 1: Description of focus group members.

Member	Position	Age Range
1	Product Designer and PhD student in the area of Human-Computer Interaction	25-30
2	Master's student in the area of Human-Computer Interaction	25-30
3	PhD student in Artificial Intelligence	30-35
4	Public manager in the area of Technology and PhD candidate in the area of Systems Engineering and Smart Cities	35-40
5	Leader of Corporate Architecture and of Solutions in a public company and Master's student in the field of Artificial Intelligence	35-40

5 HIPPOLYTA, THE PROPOSED FRAMEWORK

In previous works [4], we presented that the open data manipulation flow happens in some stages. First, citizens reflect on their needs and what they want to look for in an OGDP. In a second moment, citizens collect data on the portal, through the search interface, or through the categories present in the menu. Soon after, the citizen performs a data merge to gather all relevant data to the search question. Afterward, citizens still need to clean and process the data to remove empty, null, and inconsistent fields. Later, with the data cleaned and treated, citizens must create visualizations highlighting certain aspects, such as trends and outliers. Finally, after all the steps described, citizens can conclude the research subject [4].

Furthermore, still in [4], we discovered that several aspects harm citizens through data manipulation. There are notorious problems, both in the analyzed open data portal and in data visualization tools, which permeate the current flow of open data manipulation, such as (i) much time spent searching and browsing the portal to identify sets of ideal data; (ii) non-intuitive interfaces; (iii) lack of tutorials; (iv) excessive difficulty in accessing the platforms; (v) inconsistent data and (vi) limited features.

5.1 Results

We have divided Hippolyta into three main modules: (i) the semantic enrichment module, which works as a direct textual communication interface between the citizen and the application; (ii) the data retrieval module, which collects data from OGDPs; and (iii) the data visualization module, which also works as a direct interface with the citizen, but already shows the final result of the research carried out through data visualizations.

5.1.1 Semantic enrichment module. This module represents Hippolyta's first task. In this module, the citizen interacts through a search interface, similar to the interface of significant internet search engines. We implement Natural Language Processing (NLP) techniques in the backend to identify the citizen's needs and to make an information extraction to point out the citizen's real purpose when typing his search in the application's search bar.

In order to test and identify the lessons learned, we used a classifier, trained by [8], based on a multilayer perceptron (MLP) neural network and vector space model, for the task of part-of-speech tagging. The MLP network examines words, produces a score to assign each tag to each word, and then determines the tags using the Viterbi algorithm. Also, we apply stemming processing, which is the process of deriving the basic word by removing the affix from the word [27], and lemmatization process, which is the process of determining the dictionary form of a word, given one of its inflected variants [27], as we can observe in Figure 3.

5.1.2 Data retrieval module. This module represents the second task of Hippolyta, and aims to allow data retrieval, performing a search task in an open data catalog. In order to test and identify lessons learned, we used the CKAN API or Comprehensive Knowledge Archive Network. CKAN[1] is a web-based management system developed by Open Knowledge Foundation[2], and more than 192 governments use it, institutions, and other organizations worldwide to manage open government data. CKAN, written in Python, uses Solr, an open-source Java-based information retrieval library, to achieve full-text search functionality on datasets stored in PostgreSQL backup. The CKAN API is useful for developers who want to write code interacting with CKAN websites and their data. The API is also extensively documented and provides a comprehensive way to retrieve metadata from the Data Catalog [13]. For data retrieval, it is necessary to perform an API request, where the response is a JSON file in data dictionary format. From the JSON file, we run an algorithm that examines the entire key/value structure and automatically performs the search and alignment between data, metadata, and keywords extracted in the previous module.

In previous work [4], we applied and validated this module. We used metrics to evaluate information retrieval, such as precision and recall, and obtained satisfactory results in applying the proposed methodology.

5.1.3 Data visualization module. This module represents Hippolyta's third task and performs direct communication between the citizen and the application. One of the biggest hurdles in creating data visualizations is the need for technical knowledge. With the openness of government data and the lack of standards, the need to process

[1] https://ckan.org/
[2] https://okfn.org/

data before and during creating a data visualization is explicit. Although some tools already allow citizens to process data, this action requires extra steps and a minimum knowledge of data formats that discourages a slice of society [9]. In order to test and identify lessons learned, implementing data visualizations required a few steps: (i) classification of data types; (ii) transforming the data, and (iii) deciding and building the appropriate data views. We already used all these steps in our previous work [4].

We followed a standard process in tools like Tableau[3] [16] and DeepEye[4] [21], classifying data types into categories. Moreover, we decide the visualization type through classified data types and build the visualizations through python libraries such as *Seaborn*[5], *Bokeh* [6] and *Altair* [7], depending on complexity and need for each selected data visualization.

5.2 Hippolyta instantiation

In order to materialize Hippolyta so that the framework could process and visualize the available data, we used Streamlit [8]. Streamlit is an open-source Python library used to build web applications for machine learning and data science. Streamlit simplifies the job of coding and viewing results in the web application. After implementing Hippolyta through Streamlit, we were able to use its features and evaluate whether the framework is capable of helping citizens to interpret open government data. Reinforcing that, we developed Hippolyta's instantiation in Portuguese, so for this publication, we edited the images for the English language. In Figure 2, we can see the initial screen of Hippolyta, where the citizen writes his need and the semantic enrichment module is executed. We emphasize that the full framework instantiation was carried out in Portuguese to evaluate it with Brazilian citizens.

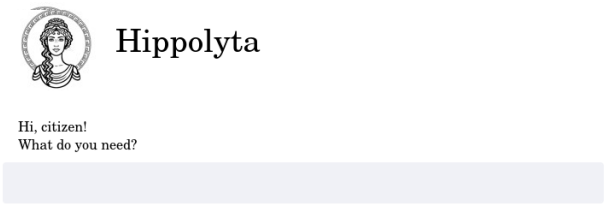

Figure 2: Hippolyta home screen.

In Figure 3, after writing "I want to know about covid-19", Hippolyta already executed the semantic enrichment module, and we observe that the data retrieval module is being executed to deliver available datasets to the user.

In Figure 4, after the citizen selects which fields he wants to visualize, in this case about "contracts coronavirus - covid 19 - April 2020", related to index number 14, the data visualization module is executed, and we have the results delivered to the citizen.

[3]https://www.tableau.com/pt-br
[4]http://deepeye.tech/
[5]https://seaborn.pydata.org/
[6]https://docs.bokeh.org/en/latest/index.html
[7]https://altair-viz.github.io/
[8]https://streamlit.io/

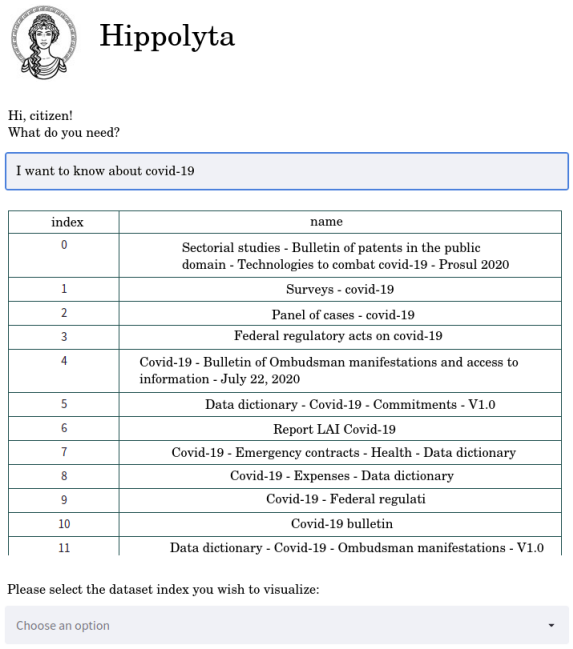

Figure 3: Hippolyta data retrieval.

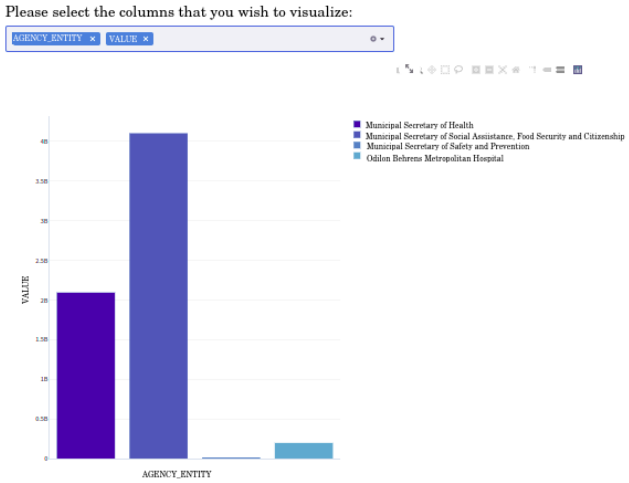

Figure 4: Interactive data visualization in Hippolyta.

We can obeserve another example of using Hippolyta in Figure 5. However, as the data visualization module can classify the data type, transform it, and decide the appropriate visualization, in this example, the visualization selected by Hippolyta was the scatter plot.

6 INTERPRETABILITY EVALUATION

First, data interpretability was presented to members in the context of open government data and its characteristics. After clarifying

Figure 5: Use of Hippolyta for a scenario where citizens want to know more about UFF.

the concept, we used usage scenarios in Hippolyta and the Brazilian Open Data Portal. Considering that we considered the guiding question "Does the Hippolyta Framework provide more interpretability to the citizen?" for conducting the focus group, a broader discussion was centered on eight main questions (MQ). Because of this, for each MQ, the focus group reached a consensus, obtaining unified comments. The responses are summarized below:

[MQ1] How do you feel about the presentation of the data (understandable, simple, clear, readable) when comparing the use of Hippolyta with the use of the Portal?

- "Hippolyta's existing data visualization functionality really facilitates data interpretation. Hippolyta presents the data in a way that consumption is pleasant. The possibility of a semantic search also facilitates and simplifies the search for information. The existence of the semantic search provides greater comfort for those who are searching for information. The Brazilian Open Data Portal only brings raw data, in which the interpretation depends on a larger and more complex process of data cleaning and analysis".

The focus group continually emphasized that the data visualization module is a focal point to leverage data interpretability, as the data is already clean and in an appropriate graphical form, different from that presented by the Brazilian portal.

[MQ2] What is your feeling about the data reliability and traceability when we compare the use of Hippolyta with the Portal?

- "We noticed a lack of data traceability in Hippolyta, which can directly impact questions about reliability. As the Brazilian Open Data Portal is the direct responsibility of the Federal Government, citizens tend to rely more on available data".

The focus group highlighted that citizens would trust the Brazilian open data portal more due to the official nature of the platform. However, Hippolyta depends directly on the data sources of the Brazilian portal.

[MQ3] What is your feeling about data structuring and organization when we compare the use of Hippolyta with the Portal?

- "Hippolyta's simple and direct layout, compared to the layout of the Brazilian Open Data Portal, helps to avoid confusing citizens when data searching".

The focus group emphasized the importance of maintaining a more organized and minimalist structure to make the citizens' experience less confusing.

[MQ4] What is your feeling about the data accuracy and correctness when we compare the use of Hippolyta with the Portal?

- "As Hippolyta directly depends on the source of data origin and does not clearly show where this data is collected, questions about correctness and precision can be impacted since the Brazilian Open Data Portal, which is responsible for making available accurate and correct data".

As previously mentioned, Hippolyta depends directly on the data sources of the Brazilian portal. Therefore, the accuracy and correctness of the data also depend on this data source.

[MQ5] What is your feeling about the data completeness when we compare the use of Hippolyta with the Portal?

- "As Hippolyta does not provide further descriptions of the collected data, questions about data completeness may be negatively impacted. However, the Brazilian Open Data Portal also needs to improve in this characteristic."

The focus group pointed out that Hippolyta, like the portal, no longer offers descriptions of the datasets. However, if the source does, Hippolyta can easily be extended to provide descriptions of the datasets.

[MQ6] How do you feel about the data conciseness when we compare the use of Hippolyta with the Portal?

- "Hippolyta provides better data conciseness, considering that in addition to the framework treating the data in a way that the citizen does not need to worry about duplicates. Also, the citizen can directly select the relevant variables for his search for information, excluding thus unnecessary or irrelevant data. The Brazilian Open Data Portal provides duplicate data without selecting what would be relevant or not to the citizen."

Since Hippolyta already offers pre-cleaning of data, such as removing duplicates, conciseness becomes a feature understood by the framework.

[MQ7] What is your feeling about the data consistency and coherence when we compare the use of Hippolyta with the Portal?

- "As the data origin directly impacts Hippolyta, questions about data integrity and logic are also directly impacted."

Making Hippolyta more independent of the original data source would bring us a significant gain because, again, the data source becomes a weakness.

[MQ8] What is your feeling about the data informativeness when we compare the use of Hippolyta with the Portal?

- "Hippolyta works in a more direct, clearer, simpler way and consequently more attractive. The possibility for the user to directly type their need is closer to everyday reality. On the other hand, the Brazilian Open Data Portal has a confusing layout and needs an improved search, alienating the citizen."

Hippolyta is more attractive to citizens due to its simplicity, the semantic enrichment module, which brings citizens greater comfort in researching their needs. Nevertheless, also for the other modules that facilitate the user's cognitive effort. We also considered four secondary questions (SQ). Given this, the focus group reached a consensus for each SQ, obtaining unified comments. The responses are summarized below:

[SQ1] Comment on Hippolyta's potential.

- "With a different form of interaction, Hippolyta can help in greater citizen engagement."
- "The existence of a search closer to natural language can help bring citizens closer to the process of consuming open government data."
- "Rapidly translating the user question into data visualization, in a simple way, minimizes the time of collection, cleaning, and insights generation."
- "Hippolyta can help citizens to develop a new habit and expands the possibility of direct intervention by citizens in decision-making procedures and control of the exercise of power."
- "Hippolyta can also be used as a teaching tool in popular classes and school education."

[SQ2] Comment on Hippolyta's weaknesses.

- "Hippolyta's lack of auditability is a visible weakness, but one that can be improved in the short term."

[SQ3] Comment on Hippolyta's opportunities.

- "Hippolyta brings an opportunity to change paradigms in participatory democracy, intending to change interaction, closer to the citizen."

[SQ4] Comment on Hippolyta's challenges.

- "Hippolyta's great challenge is that it is a data consumer directly impacted by the source data quality."

Table 2 shows some evolutions proposed by the focus group to make Hippolyta's functionalities more complete.

According to the focus group, the groups in which Hippolyta is robust are (i) G1 — Comprehensibility, simplicity, clarity, and readability, (ii) G3 — Structuring and organization, (iii) G6 — Conciseness and (iv) G8 — Informativeness. Also, according to the focus group, the groups in which Hippolyta presents vulnerabilities are (i) G2 — Reliability and traceability, (ii) G4 — Precision and correctness, (iii) G5 — Completeness and (iv) G7 — Consistency and coherence.

The focus group claimed that G2 — Reliability, and traceability is compromised due to the lack of some data traceability. And then, as the Brazilian Open Data Portal is the direct responsibility of the Federal Government, citizens tend to trust the available data more. However, a data traceability resource can be easily implemented in Hippolyta with direct links to the origin of the retrieved data.

Table 2: Hippolyta's evolutions, proposed by the focus group, given each characteristics group that constitutes the concept of interpretability.

Characteristics Group	Proposed Improvements
G1	Inclusion of a tutorial to help citizens understand the rendered visualization.
G2	Inclusion of direct links to the retrieved data source.
G3	-
G4	Inclusion of a Citizensourcing process for inserting validations about the correctness of retrieved data.
G5	Inclusion of a citizen-sourcing process to insert more complete data descriptions.
G6	Inclusion of the citizen-sourcing process in order to expose the real relevance of the collected data.
G7	Inclusion of feedback feature like: "Was your question answered?"
G8	-

About the other groups (G4, G5, and G7), considering that Hippolyta is directly impacted by the origin of the data, in our case by the Brazilian Open Data Portal, which proves that the groups in question also have weaknesses in the portal in question, we raise proposals for evolution so that Hippolyta can reduce this impact. As indicated in Table 2, the inclusion of a citizen-sourcing process in Hippolyta can help considerably to improve the respective issues — accuracy, correctness, completeness, consistency, and coherence of the data — positioning citizens as data auditors as well as users.

Also, in previous works [4], we noticed that the most significant difficulties of the Brazilian Open Data Portal are related to groups G1, G3, and G8, groups in which Hippolyta has some competence. An example that reinforces the fragility of the Brazilian Open Data Portal, which ends up impacting the use of Hippolyta, is the high rate of unavailability of the portal, which we sometimes faced in the implementation process of Hippolyta.

7 CONCLUSIONS

The general objective of this work is to propose the definition, implementation, and evaluation of a framework named Hippolyta, which is capable of helping citizens to interpret open government data.

Therefore, we defined and developed the architecture of a framework named Hippolyta, which (i) identifies the citizen's needs using a semantic enrichment module, (ii) performs the data collection through a data retrieval module, (iii) creates a visualization from the dataset chosen by the citizen, through a data visualization module. So, to understand whether Hippolyta could help citizens to interpret the available data, we carried out a set of evaluations. Considering the first evaluation performed on Hippolyta, the results were satisfying, considering that high values of precision and recall indicate a high data recovery power of the framework. The second evaluation of Hippolyta consolidated her ability to improve interpretability, considering that the best-evaluated functionalities

of Hippolyta represent fundamental caractheristics for the interpretability definition.

To evaluate whether Hippolyta could boost the interpretability of open government data, we carried out a qualitative research method called a focus group, in which the members interacted with Hippolyta. As a result of the sessions with the focus group, we obtained comments on how Hippolyta evidence, or not, each group of characteristics — groups that constitute the data interpretability definition. The focus group also proposed evolutions for Hippolyta, considering its functionalities. In summary, we verified Hippolyta's competence in promoting the interpretability of open government data.

The scientific contributions of this study are as follows. While various open data studies have indicated that open government data is often difficult to interpret or that there is a risk of open data misinterpretation [7, 18, 30], the open data literature lacks insight into how open data interpretability can be enhanced. This study contributes to the open data literature by developing a framework that provides in-depth insights into what elements the open data interpretability concept is composed of and how different modules of the framework can enhance open data interpretability. The main practical and societal contribution of this study is that policymakers and other professionals can apply our Hippolyta framework to promote greater interpretability of open government data. Consequently, this can potentially reduce the social and technical barriers imposed by the difficulty of citizens to correctly interpret the data made available in the open government data projects. Subsequently, the correct interpretation of OGD by citizens potentially leads to more value creation with open data, including social, economic, and operational value.

We consider the main limitation of our work that different profiles of potential users present in society should have evaluated Hippolyta. For future research, the interactive design technique based on continuous evaluation [12] may help us to cover different citizen profiles.

REFERENCES

[1] Gabriel Abu-Tayeh, Oliver Neumann, and Matthias Stuermer. 2018. Exploring the motives of citizen reporting engagement: Self-concern and other-orientation. *Business & information systems engineering* 60, 3 (2018), 215–226.

[2] Dani Arribas-Bel, Mark Green, Francisco Rowe, and Alex Singleton. 2021. Open data products-A framework for creating valuable analysis ready data. *Journal of Geographical Systems* 23, 4 (2021), 497–514.

[3] Judie Attard, Fabrizio Orlandi, and Sören Auer. 2016. Value creation on open government data. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2605–2614.

[4] Raissa Barcellos, Flavia Bernardini, and José Viterbo. 2022. Towards defining data interpretability in open data portals: Challenges and research opportunities. *Information Systems* 106 (2022), 101961.

[5] Iván Cantador, Jesús Viejo-Tardío, María E Cortés-Cediel, and Manuel Pedro Rodríguez Bolívar. 2021. A Chatbot for Searching and Exploring Open Data: Implementation and Evaluation in E-Government. In *DG. O2021: The 22nd Annual International Conference on Digital Government Research*. 168–179.

[6] Abiola Paterne Chokki, Benoît Frénay, and Benoît Vanderose. 2022. Open Data Explorer: An End-to-end Tool for Data Storytelling using Open Data. In *AMCIS 2022*.

[7] Peter Conradie and Sunil Choenni. 2014. On the barriers for local government releasing open data. *Government information quarterly* 31 (2014), S10–S17.

[8] Erick Fonseca and João Luís G Rosa. 2013. Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian symposium in information and human language technology*.

[9] Alvaro Graves and James Hendler. 2013. Visualization tools for open government data. In *Proceedings of the 14th Annual International Conference on Digital Government Research*. 136–145.

[10] Ruben Interian, Isela Mendoza, Flavia Bernardini, and José Viterbo. 2022. Unified vocabulary in Official Gazettes: An exploratory study on procurement data. In *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*. 195–202.

[11] Ellen Johnson. 2021. Face validity. In *Encyclopedia of autism spectrum disorders*. Springer, 1957–1957.

[12] Sabine Junginger. 2016. *Transforming Public Services by Design: Re-orienting policies, organizations and services around people*. Taylor & Francis.

[13] Fabian Kirstein, Benjamin Dittwald, Simon Dutkowski, Yury Glikman, Sonja Schimmler, and Manfred Hauswirth. 2019. Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP. In *International Conference on Electronic Government*. Springer, 192–204.

[14] Julia Lanoue. 2020. Disparate Environmental Monitoring as a Barrier to the Availability and Accessibility of Open Access Data on the Tidal Thames. *Publications* 8, 1 (2020), 6.

[15] Sabina Leonelli. 2019. Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review* 1, 1 (2019).

[16] Jock Mackinlay, Pat Hanrahan, and Chris Stolte. 2007. Show me: Automatic presentation for visual analysis. *IEEE transactions on visualization and computer graphics* 13, 6 (2007), 1137–1144.

[17] Gustavo Magalhaes and Catarina Roseira. 2020. Open government data and the private sector: An empirical view on business models and value creation. *Government Information Quarterly* 37, 3 (2020), 101248.

[18] Sébastien Martin. 2013. Risk analysis to overcome barriers to open data. *Electronic Journal of e-Government* 11, 2 (2013), pp348–359.

[19] Lokanath Mishra. 2016. Focus group discussion in qualitative research. *Techno Learn* 6, 1 (2016), 1.

[20] Andrés Moreno, José Molano-Pulido, Juan E Gomez-Morantes, and Rafael A Gonzalez. 2022. ADACOP: A Big Data Platform for Open Government Data. In *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*. 369–375.

[21] Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2018. DeepEye: An automatic big data visualization framework. *Big data mining and analytics* 1, 1 (2018), 75–82.

[22] Pamela Robinson and Teresa Scassa. 2022. *The Future of Open Data*.

[23] A Paula Rodriguez Müller. 2020. Making smart cities "smarter" through ICT-enabled citizen coproduction. *Handbook of smart cities* (2020), 1–21.

[24] Erna HJM Ruijer and Evelijn Martinius. 2017. Researching the democratic impact of open government data: A systematic literature review. *Information Polity* 22, 4 (2017), 233–250.

[25] Igor Garcia Ballhausen Sampaio, Eduardo de O Andrade, Flávia Bernardini, and José Viterbo. 2022. Assessing the Quality of Covid-19 Open Data Portals. In *International Conference on Electronic Government*. Springer, 212–227.

[26] Abu-El Seoud, Ralf Klischewski, et al. 2015. Mediating citizen-sourcing of open government applications–a design science approach. In *International Conference on Electronic Government*. Springer, 118–129.

[27] Ayush Srivastav, Hera Khan, and Amit Kumar Mishra. 2020. Advances in Computational Linguistics and Text Processing Frameworks. In *Handbook of Research on Engineering Innovations and Technology Management in Organizations*. IGI Global, 217–244.

[28] Sue P Stafford. 2009. Data, information, knowledge, and wisdom. *Knowledge Management, Organizational Intelligence and Learning, And Complexity* 3 (2009), 179.

[29] Xiaohua Zhu and Mark Antony Freeman. 2019. An evaluation of US municipal open data portals: A user interaction framework. *Journal of the Association for Information Science and Technology* 70, 1 (2019), 27–37.

[30] Anneke Zuiderwijk and Marijn Janssen. 2014. Barriers and development directions for the publication and usage of open data: A socio-technical view. In *Open government*. Springer, 115–135.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009